

Kernel Density Estimation with Berkson Error

James P. Long

Department of Statistics, Texas A&M University
3143 TAMU, College Station, TX 77843-3143
jlong@stat.tamu.edu

Noureddine El Karoui

Department of Statistics, University of California, Berkeley
367 Evans Hall # 3860, Berkeley, CA 94720-3860
nkaroui@stat.berkeley.edu

John A. Rice

Department of Statistics, University of California, Berkeley
367 Evans Hall # 3860, Berkeley, CA 94720-3860
rice@stat.berkeley.edu

Abstract

Given a sample $\{X_i\}_{i=1}^n$ from f_X , we construct kernel density estimators for f_Y , the convolution of f_X with a known error density f_ϵ . This problem is known as density estimation with Berkson error and has applications in epidemiology and astronomy. Little is understood about bandwidth selection for Berkson density estimation. We compare three approaches to selecting the bandwidth both asymptotically, using large sample approximations to the MISE, and at finite samples, using simulations. Our results highlight the relationship between the structure of the error f_ϵ and the optimal bandwidth. In particular, the results demonstrate the importance of smoothing when the error term f_ϵ is concentrated near 0. We propose a data-driven bandwidth estimator and test its performance on NO₂ exposure data.

Keywords: Berkson Error; Measurement Error; Bandwidth Selection; Kernel Density Estimation; Multivariate Density Estimation

Short title: Kernel Density Estimation with Berkson Error

1 Introduction

1.1 Background

We consider smoothing a density estimate when an error-free sample is observed and one is interested in the convolution of the population density with an error term. This is known as density estimation with Berkson error and has been studied in Delaigle [2007] where NO_2 exposure in children is estimated using known kitchen and bedroom concentrations. The exposure level in children is modeled as a function of kitchen and bedroom concentrations plus independent, additive random error.

Density estimation with Berkson error is one example of a class of problems where low-error or error-free data is used to construct an estimate for noisy data. Bovy et al. [2011] considers this problem in the context of constructing a classifier for noisy astronomical data. Here, each object belongs to the class quasar or star. For each object a telescope records a vector of flux ratios. For a training set of observations of known class, the authors observe low-error flux ratios. However for the data of unknown class, flux ratios contain significant measurement error. The goal is to construct an accurate classifier for the noisy data.¹ Carroll et al. [2009] considers a similar problem in a regression context with data arising from nutritional epidemiology. Long et al. [2012] studies this problem in the context of classification of periodic variable stars.

In each of these works, tuning parameters are selected to optimize some risk function. There is extensive literature on selecting tuning parameters for problems where all data is observed without measurement error. For example with kernel density estimation, asymptotic rates for the mean integrated squared error (MISE) as a function of bandwidth as well as finite-sample procedures for selecting the bandwidth are known [Jones et al., 1996, Silverman, 1986]. In contrast, much less is understood about selection of tuning parameters when one set of data is error-free and there is measurement error in the variable of interest. In this work, we derive asymptotic results and present finite sample simulations illustrating the relationship between measurement error and smoothing for kernel density estimators. While our results are most directly applicable to density estimation with Berkson error, they have implications for the classification and regression problems above.

We now formalize the density estimation problem. Suppose we observe independent $\{X_i\}_{i=1}^n \sim f_X$. We seek to use this data to estimate the density, denoted f_Y , of $Y = X + \epsilon$. Here $\epsilon \sim f_\epsilon$, $X \sim f_X$, and ϵ and X are independent. All random variables are in \mathbb{R}^p . In the literature, ϵ is known as Berkson error and was introduced in a regression context by Berkson [1950]. It differs from classical measurement error where one

¹See Section 2 (Equations 1, 2, and 3) and Section 5 of Bovy et al. [2011] for more information.

observes an error contaminated sample and seeks to estimate the underlying, uncontaminated density.²

For estimating f_Y , Delaigle [2007] proposed using

$$\tilde{f}_Y(y) = \frac{1}{n} \sum_{i=1}^n f_\epsilon(y - X_i). \quad (1)$$

Delaigle [2007] showed that when f_ϵ is square-integrable and f_X is bounded, this estimator is unbiased with a mean integrated squared error (MISE) that converges to 0 at rate n . The convergence result contrasts with standard density estimation where the MISE is generally of order $n^{-4/(4+p)}$.³ Effectively, knowledge of f_ϵ provides valuable information about the structure of f_Y unavailable in the standard kernel density estimation case. In addition to a fast convergence rate, \tilde{f}_Y in Equation 1 has no tuning parameters that require estimation. A potential drawback of \tilde{f}_Y in Equation 1 is that for cases where ϵ is concentrated around 0, the estimator will have large spikes at the sample points $\{X_i\}_{i=1}^n$ and thus high variance. We illustrate this problem through simulations in Section 4.

In this work we study of impact of smoothing estimates of f_Y using kernels. We focus on comparing three approaches to kernel bandwidth selection:

- **Approach 1:** Select a bandwidth specifically to optimize estimation of f_Y .
- **Approach 2:** Since $f_Y = \int f_X(y - \epsilon)f_\epsilon(\epsilon)d\epsilon$, use a kernel density estimator to estimate f_X . Select the bandwidth to optimize estimation of f_X . Then convolve this estimate of f_X with f_ϵ in order to estimate f_Y .
- **Approach 3:** Set the bandwidth to 0 i.e., use Delaigle’s estimator in Equation (1).

Each of these approaches has some attractive properties. Approach 1 may provide the optimal performance because the bandwidth is chosen specifically to estimate f_Y . Approach 2 is attractive because there is extensive literature on selecting a bandwidth which optimizes estimation of f_X . Approach 3 is attractive because there are no tuning parameters to estimate and the resulting estimator has parametric first order convergence rates as shown by Delaigle [2007].

In addition to shedding light on the Berkson density estimation problem, a comparison of the performance of these approaches may be useful when considering how to regularize classification or regression methods when training data is error-free (or low-error) and data of unknown class has measurement error.

²See Carroll et al. [2006] for a review of Berkson and classical measurement error.

³The $n^{-4/(4+p)}$ order for the MISE requires regularity conditions on f_Y . For example, Wand and Jones [1995] (Section 4.3, p.95) assumes each entry of the Hessian of f_Y is piecewise continuous and square integrable. See p.100 of Wand and Jones [1995] for the MISE convergence rate.

1.2 Outline of Work and Summary of Findings

In Section 2 we present a kernel bandwidth estimator for f_Y that allows simultaneous comparison of all three smoothing approaches. As a guide towards comparing the approaches, in Section 3 we derive a second order expansion of the MISE of the estimator as a function of bandwidth. The asymptotic expansion reveals interesting properties:

- Asymptotically, the optimal bandwidth for estimating f_Y (approach 1) is of order $n^{-1/2}$. Notably this rate does not depend on the dimension of the problem, unlike for standard kernel density estimation. Approach 1 provides a second order (n^{-2}) reduction in MISE over approach 3 (no smoothing).
- The asymptotically optimal bandwidth for estimating f_Y depends on f_X which is unknown. The expression we derive for the optimal bandwidth is used as the basis for a plug-in estimator in Section 5.
- Using approach 2 (smoothing to optimize estimation of f_X) results in a MISE of order $n^{-4/(4+p)}$, slower than the n^{-1} order of approaches 1 and 3. As the dimension increases, the discrepancy in these rates grows.

Based purely on asymptotic considerations, approaches 1 and 3 appear superior to approach 2. In Section 4 we study the finite sample properties of these three approaches by adapting a result from Wand and Jones [1993] which allows for exact computation of MISE when f_X is a mixture of normals and the error and kernel are normal. We find:

- Approach 3 (no smoothing) can result in drastically undersmoothing the density, particularly when the error term is concentrated around 0.
- Approach 2 (smoothing to optimize estimation of f_X) can result in oversmoothing the density, particularly when the error term is smooth 3.
- In the cases considered in the simulation, the qualitative impacts of undersmoothing by using approach 3 appear worse than the qualitative impacts of oversmoothing by using approach 2.
- Approach 1 outperforms approach 2 or 3. In the simulations considered, this performance advantage increased in three dimensions versus one dimension.

In Section 5 we propose a data-based bandwidth estimator for approach 1 and apply our methodology to the NO₂ data studied by Delaigle [2007]. In Section 6 suggest some directions for future work. Proofs of all theorems are given in Section S.1 and some technical issues are addressed in Section S.2.

2 Problem Setup

We observe independent random variables $X_1, \dots, X_n \sim f_X$. We aim to estimate, f_Y , the density of

$$Y = X + \epsilon.$$

Here $X \sim f_X$, $\epsilon \sim f_\epsilon$, and X and ϵ are independent. f_ϵ is assumed known. All random variables are in \mathbb{R}^p . In all that follows let \hat{f}_V represent the characteristic function of the random variable V and let \tilde{f} represent an estimator of f .

2.1 Construction of an Estimator for f_Y

We construct an estimator for f_Y by first estimating \hat{f}_Y , the characteristic function of f_Y . Let K be a mean 0 density function called the kernel, and \hat{K} its characteristic function. Let

$$\Sigma_K = \int xx^T K(x) dx.$$

Let $H = H_n \succeq 0$ be a sequence of positive semidefinite $p \times p$ matrices called the bandwidth.

$$\tilde{f}_X(\omega) = \frac{1}{n} \sum_{j=1}^n e^{i\omega^T X_j}$$

is an estimate of \hat{f}_X . Consider estimating \hat{f}_Y (the characteristic function of f_Y) using

$$\tilde{f}_{Y,H}(\omega) = \hat{K}(H\omega) \hat{f}_\epsilon(\omega) \tilde{f}_X(\omega). \quad (2)$$

Note that $\tilde{f}_{Y,H}$ is a characteristic function because it is the product of characteristic functions. Assuming $\tilde{f}_{Y,H} \in L_1$, we may estimate f_Y using

$$\tilde{f}_{Y,H}(y) = \frac{1}{(2\pi)^p} \int e^{-i\omega^T y} \tilde{f}_{Y,H}(\omega) d\omega. \quad (3)$$

The assumption that $\tilde{f}_{Y,H} \in L_1$ implies $\tilde{f}_{Y,H}$ is a bounded density (see Theorem 3.3 in Durrett [2005]). Throughout this work, we require $\hat{f}_\epsilon \in L_1$, thus guaranteeing that $\tilde{f}_{Y,H} \in L_1$ and ensuring that $\tilde{f}_{Y,H}$ in Equation (3) is a valid density.

2.2 $\tilde{f}_{Y,H}$ and Approaches to Smoothing

$\tilde{f}_{Y,H}$ allows for simultaneous comparison of all three bandwidth selection approaches discussed in the introduction. With approach 1, we optimize H in $\tilde{f}_{Y,H}$ specifically for estimating f_Y . Recall that with approach 2, we construct a kernel density estimator for f_X with some bandwidth $H_X \succ 0$ optimized for estimating f_X . Denote this estimator

$$\tilde{f}_{X,H_X}(x) = \frac{1}{n} \sum_{i=1}^n K_{H_X}(x - X_i).$$

Since $f_Y(\epsilon) = \int f_X(y - \epsilon) f_\epsilon(\epsilon) d\epsilon$, $\tilde{f}_{X,H_X}(x)$ is convolved with f_ϵ to estimate f_Y . However this procedure is equivalent to using H_X in $\tilde{f}_{Y,H}$. In other words,

$$\tilde{f}_{Y,H_X}(y) = \int \tilde{f}_{X,H_X}(y - \epsilon) f_\epsilon(\epsilon) d\epsilon.$$

Finally $\tilde{f}_{Y,H}$ includes as a subcase approach 3, no smoothing. By setting $H = 0$ we have

$$\tilde{f}_{Y,0}(y) = \frac{1}{n} \sum_{i=1}^n f_\epsilon(y - X_i).$$

This is the kernel-free estimator of Delaigle [2007] presented in Equation (1).

2.3 MISE

We use mean integrated squared error as a guide toward understanding the behavior of the three approaches to selecting the bandwidth in $\tilde{f}_{Y,H}$. Let \mathcal{P}_n be the product measure on (X_1, \dots, X_n) . The mean integrated squared error is defined as

$$\text{MISE}(H) \equiv \mathbf{E}_{\mathcal{P}_n} \int \left(\tilde{f}_{Y,H}(y) - f_Y(y) \right)^2 dy.$$

MISE is a popular measure of risk used in many density estimation studies (see for example Delaigle [2008], Jones et al. [1996], Marron and Wand [1992], Tsybakov [2009], Wand and Jones [1995]). The MISE allows for relatively straightforward asymptotic analysis (Theorems 1 and 2) as well as admitting to, under certain conditions, a computationally efficient representation which we exploit in order to obtain our finite sample results (Theorem 3). We explore some of the qualitative impacts of the different smoothing approaches, not directly captured by MISE, in Subsection 4.3. While other measures, such as K-L divergence, Hellinger distance, or integrated absolute error, could be used, we feel MISE captures the essential properties of the three approaches to smoothing.

The optimal H , in terms of minimizing MISE for estimating f_Y , is

$$H_Y = \underset{\{H: H \geq 0\}}{\text{argmin}} \text{MISE}(H).$$

This is the bandwidth of approach 1. For approach 2, the bandwidth is chosen to minimize the MISE in estimating f_X . In other words

$$H_X = \operatorname{argmin}_{\{H: H > 0\}} \mathbf{E}_{\mathcal{P}_n} \int \left(\tilde{f}_X(x) - f_X(x) \right)^2 dx.$$

Unfortunately the MISE expression is complicated and exact expressions for H_Y and H_X are not generally possible. In Section 3 we form asymptotic approximations to the MISE and determine the rates at which $\|H_Y\|_\infty \rightarrow 0$ and $\text{MISE}(H_Y) \rightarrow 0$ as $n \rightarrow \infty$. We refer to standard kernel density theory for the rates associated with H_X . Our asymptotic expansion of the MISE reveals rates associated with $\text{MISE}(H_X)$. The asymptotic approximation also shows the error for approach 3, $\text{MISE}(0)$, a quantity that was derived by Delaigle [2007].

In Section 4 we specialize to the case where f_X is a Gaussian mixture, K is a Gaussian kernel, and f_ϵ is a Gaussian error density. In this setting, the MISE can be evaluated at a particular H without numerically approximating integrals (see Theorem 3). Using this result we study the finite sample properties of H_Y , H_X , $\text{MISE}(H_Y)$, $\text{MISE}(H_X)$, and $\text{MISE}(0)$.

In this work, H_Y refers to the exact optimal bandwidth, H_Y^* refers to an asymptotically optimal bandwidth, and \tilde{H}_Y refers to a data based estimator of H_Y (used mostly in Section 5). While we derive asymptotic results for bandwidth matrices, we often specialize to the case of a scalar bandwidth. In such cases, lowercase letters, h_Y , h_Y^* , and \tilde{h}_Y , are used. The same notation applies for H_X .

3 Asymptotic Results

For the purposes of forming asymptotic expansions, we represent the MISE in terms of characteristic functions.

Theorem 1. Assume $\hat{f}_Y, \tilde{f}_{Y,H} \in L_1$. Then

$$(2\pi)^p \text{MISE}(H) = \int |1 - \hat{K}(H\omega)|^2 d\mu(\omega) + \frac{1}{n} \int |\hat{K}(H\omega)|^2 d\nu(\omega) \quad (4)$$

where

$$\begin{aligned} d\mu(\omega) &= |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)|^2 d\omega, \\ d\nu(\omega) &= |\hat{f}_\epsilon(\omega)|^2 (1 - |\hat{f}_X(\omega)|^2) d\omega \end{aligned}$$

are positive measures.

See Subsection S.1.1 on p.S.1 for a proof. The representation of the MISE in Equation (4) closely resembles that of Tsybakov [2009] Theorem 1.4. In Equation (4), $\int |1 - \widehat{K}(H\omega)|^2 d\mu(\omega)$ is the integrated squared bias of $\widetilde{f}_{Y,H}$ and $n^{-1} \int |\widehat{K}(H\omega)|^2 d\nu(\omega)$ is the integrated variance. Notice that for fixed H , the variance decreases at rate n^{-1} while the bias is constant. When $H = 0$, $\widehat{K}(H\omega) = 1$, so the integrated squared bias term vanishes.

We require assumptions on the kernel K and the bandwidth matrix H .

Assumptions A.

$$K \text{ is a symmetric density} \tag{5}$$

$$\widehat{K} \text{ is four times continuously differentiable} \tag{6}$$

$$H = H_n \succeq 0 \text{ (i.e. sequence is positive semidefinite)} \tag{7}$$

$$\|H\|_\infty \rightarrow 0 \tag{8}$$

Since we choose the kernel and bandwidth matrix, these assumptions can always be satisfied in practice. Common kernel choices such as the standard normal and uniform on $[-1, 1]^p$ satisfy Assumptions 5 and 6. In Assumption 7, notice that the bandwidth does not have to be strictly positive definite, unlike in the standard kernel density estimation case. We require the following assumptions on the characteristic functions of f_X and f_ϵ .

Assumptions B.

$$\int \|\omega\|_\infty^8 |\widehat{f}_\epsilon(\omega)|^2 d\omega < \infty \tag{9}$$

$$\int |\widehat{f}_\epsilon(\omega)| d\omega < \infty \tag{10}$$

Assumptions 9 and 10 are satisfied as long as the error term has a density that is smooth, such as multivariate normal or Student's t (see Sutradhar [1986] for the characteristic function of the multivariate Student's t).

Theorem 2. *Under Assumptions A and B and with the notation of Theorem 1*

$$\begin{aligned} & (2\pi)^p \text{MISE}(H) \\ &= \frac{1}{n} \int d\nu(\omega) \\ &+ \left(\frac{1}{4} \int (\omega^T H^T \Sigma_K H \omega)^2 d\mu(\omega) - \frac{1}{n} \int (\omega^T H^T \Sigma_K H \omega) d\nu(\omega) \right) (1 + O(\|H\|_\infty^2)). \end{aligned} \tag{11}$$

See Subsection S.1.2 on p.S.2 for a proof. The $n^{-1} \int d\nu(\omega)$ term is variance in the estimator that does not depend on the bandwidth. If the bandwidth is set to 0 (see approach 3 below), all other terms vanish and this is the MISE. The $\frac{1}{4} \int (\omega^T H^T \Sigma_K H \omega)^2 d\mu(\omega)$ term is bias caused by using a kernel with bandwidth H while $-n^{-1} \int (\omega^T H^T \Sigma_K H \omega) d\nu(\omega)$ is the corresponding reduction in variance.

While the full bandwidth matrix offers the most flexibility and greatest potential for reduction in MISE, this expression is difficult to optimize, see Subsection S.2.1. More simply, one could use a diagonal bandwidth matrix and optimize p bandwidths. We study this case in Subsection S.2.2. Here we focus on using a scalar bandwidth. This allows for the most direct and straightforward comparisons of the different approaches to smoothing.

3.1 Scalar Bandwidth

We reparameterize the bandwidth $H = hI$. Here the general MISE expression in Equation (11) becomes

$$\begin{aligned} (2\pi)^p \text{MISE}(h) \\ = \frac{1}{n} \int d\nu(\omega) + \left(\frac{h^4}{4} \int (\omega^T \Sigma_K \omega)^2 d\mu(\omega) - \frac{h^2}{n} \int (\omega^T \Sigma_K \omega) d\nu(\omega) \right) (1 + O(h^2)). \end{aligned} \quad (12)$$

We now discuss the three smoothing approaches

Approach 1: It is straightforward to find the bandwidth that minimizes the main terms in this MISE expression. Specifically,

$$\begin{aligned} h_Y^* &= \underset{h \geq 0}{\text{argmin}} \left(\frac{h^4}{4} \int (\omega^T \Sigma_K \omega)^2 d\mu(\omega) - \frac{h^2}{n} \int (\omega^T \Sigma_K \omega) d\nu(\omega) \right) \\ &= \sqrt{\frac{2 \int (\omega^T \Sigma_K \omega) d\nu(\omega)}{n \int (\omega^T \Sigma_K \omega)^2 d\mu(\omega)}}. \end{aligned} \quad (13)$$

h_Y^* is of order $n^{-1/2}$. This order does not depend on p , unlike for the error free case where, under some regularity conditions, the order for the asymptotically optimal amount of smoothing is $n^{-1/(4+p)}$. However, the dimension p will affect the constant on h_Y^* in Equation (13). We explore the relationship between the dimension p and the optimal amount of smoothing for finite samples in Section 4. Using, h_Y^* the MISE is

$$(2\pi)^p \text{MISE}(h_Y^*) = \frac{1}{n} \int d\nu(\omega) - \frac{1}{n^2} \frac{(\int (\omega^T \Sigma_K \omega) d\nu(\omega))^2}{(\int (\omega^T \Sigma_K \omega)^2 d\mu(\omega))} + O(n^{-3}).$$

Approach 2: Set h to minimize MISE in estimating f_X . Under certain regularity conditions on f_X , the bandwidth is order $n^{-1/(4+p)}$ (e.g. see Wand and Jones [1995] page 100). Specifically, suppose

$$h_X^* = D(n) n^{-1/(4+p)},$$

where $D : \mathbb{Z}^+ \rightarrow \mathbb{R}^+$ such that $\limsup_n D(n) < \infty$ and $\liminf_n D(n) > 0$. The MISE for estimating f_Y using h_X^* (obtained from Equation (12)) is

$$\begin{aligned} (2\pi)^p \text{MISE}(h_X^*) &= \frac{1}{n} \int d\nu(\omega) + \left(\frac{D(n)^4 n^{-4/(4+p)}}{4} \int (\omega^T \Sigma_K \omega)^2 d\mu(\omega) \right. \\ &\quad \left. - D(n)^2 n^{-(6+p)/(4+p)} \int (\omega^T \Sigma_K \omega) d\nu(\omega) \right) (1 + O(n^{-2/(4+p)})) \\ &= \left(\frac{D(n)^4 n^{-4/(4+p)}}{4} \int (\omega^T \Sigma_K \omega)^2 d\mu(\omega) \right) (1 + o(1)). \end{aligned} \quad (14)$$

The $n^{-4/(4+p)}$ order for the MISE when using h_X^* is strictly worse than the n^{-1} order that can be achieved by optimizing the bandwidth specifically for the error distribution, i.e. using h_Y^* . Essentially using h_X^* oversmooths $\tilde{f}_{Y,H}$. The first order term in $\text{MISE}(h_X^*)$, Equation (14), is caused entirely by bias.

Approach 3: Set $h = 0$. Here we have

$$(2\pi)^p \text{MISE}(0) = \frac{1}{n} \int d\nu(\omega) = \frac{1}{n} \left(\int |\hat{f}_\epsilon(\omega)|^2 d\omega - \int |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)|^2 d\omega \right).$$

Asymptotically, this approach is better than approach 2 since $\text{MISE}(0)$ is order n^{-1} . The ratio of using asymptotically optimal smoothing (h_Y^*) to no smoothing is

$$\frac{\text{MISE}(h_Y^*)}{\text{MISE}(0)} = 1 - \frac{1}{n} \frac{\left(\int (\omega^T \Sigma_K \omega) d\nu(\omega) \right)^2}{\left(\int (\omega^T \Sigma_K \omega)^2 d\mu(\omega) \right) \left(\int d\nu(\omega) \right)} + O(n^{-2}).$$

These asymptotic results suggest that using the bandwidth that minimizes error in estimating f_X for estimating f_Y is a poor idea. This procedure results in a convergence rate of higher order than either not smoothing (approach 3) or smoothing specifically for f_Y (approach 1), i.e. using h_Y^* . The asymptotic results show an improvement only at the n^{-2} level by using $h = h_Y^*$ rather than $h = 0$. Based strictly on asymptotic analysis, this small improvement in error rate may not appear to justify the extra effort required to estimate h_Y . However, simulation results in Section 4 indicate that the effects of using h_Y are more important than the asymptotic analysis suggest, both in terms of minimizing MISE and preserving important qualitative features of the densities.

4 Finite Sample Results

The asymptotic results from Section 3 illustrate the large sample behavior of the MISE under different approaches to choosing the bandwidth parameter. However it is important to understand the finite sample behavior of these quantities and what sample sizes are needed for asymptotics to be informative.

Calculation of the exact MISE in Equation 4 requires numerically approximating integrals. Therefore it is computationally challenging, given a f_X , f_ϵ , and kernel K , to determine the bandwidth H which minimizes the MISE. For the error-free kernel density estimation case, Wand and Jones [1993] showed that when f_X is a normal mixture and K is a normal kernel, the exact MISE has a simple representation that does not require numerically approximating integrals. Using this result, the authors compared bandwidth parameterizations for bivariate density estimation. Here we generalize this result to the case with Berkson measurement error. We assume f_X is a normal mixture and K and f_ϵ are normal. We use this result for studying the finite sample properties of various bandwidth selection approaches. Theorem 3 is a generalization of Theorem 1 in Wand and Jones [1993] to the case with Berkson error.

Theorem 3. *Let ϕ_Σ be the mean 0, normal density with covariance Σ . Assume the kernel $K = \phi_{\Sigma_K}$ and the error density $f_\epsilon = \phi_{\Sigma_\epsilon}$. Assume that f_X is a mixture of normal densities parameterized by $\{(\alpha_j, \mu_j, \Sigma_j)\}_{j=1}^m$ where $\sum_{j=1}^m \alpha_j = 1$ and $(\alpha_j, \mu_j, \Sigma_j)$ is the mixing proportion, mean, and variance of the j^{th} component of f_X . In other words*

$$f_X(x) = \sum_{j=1}^m \alpha_j \phi_{\Sigma_j}(x - \mu_j).$$

Let $S = H^T \Sigma_K H$. Let Ω_a for $a \in \{0, 1, 2\}$ be a $m \times m$ matrix with j, j' entry equal to

$$\phi_{aS+2\Sigma_\epsilon+\Sigma_j+\Sigma_{j'}}(\mu_j - \mu_{j'}).$$

Finally let $\alpha = (\alpha_1, \dots, \alpha_m)$. Then

$$MISE(H) = \frac{1}{n} \phi_{2S+2\Sigma_\epsilon}(0) + \alpha^T ((1 - n^{-1})\Omega_2 - 2\Omega_1 + \Omega_0)\alpha. \quad (15)$$

See Subsection S.1.3 for a proof. Equation (15) can be evaluated at a particular bandwidth H without numerically approximating integrals.

In Subsection 4.1 we compare the MISE for the three bandwidth selection approaches at finite sample sizes in one dimension. In Subsection 4.2 we repeat this analysis for several 3-dimensional densities. In Subsection 4.3 we show the visual impacts of different MISEs by plotting pointwise quantiles for density estimates using different bandwidth selection approaches. Finally in Subsection 4.4 we explore how fast the asymptotic approximations from Section 3 for the optimal smoothing parameter for f_Y take hold. All of the results presented in this section can be reproduced using publicly available code.⁴

⁴R-code and data for generating results in Sections 4 and 5 are available at <http://stat.tamu.edu/~jlong/berkson.zip>.

4.1 Relative Error in One Dimension

In this section the f_X densities are 1-dimensional normal mixtures which satisfy the assumptions of Theorem 3. The densities we consider, and associated names, are plotted in Figure 1. The exact parameter values for these densities are given in Table 1.

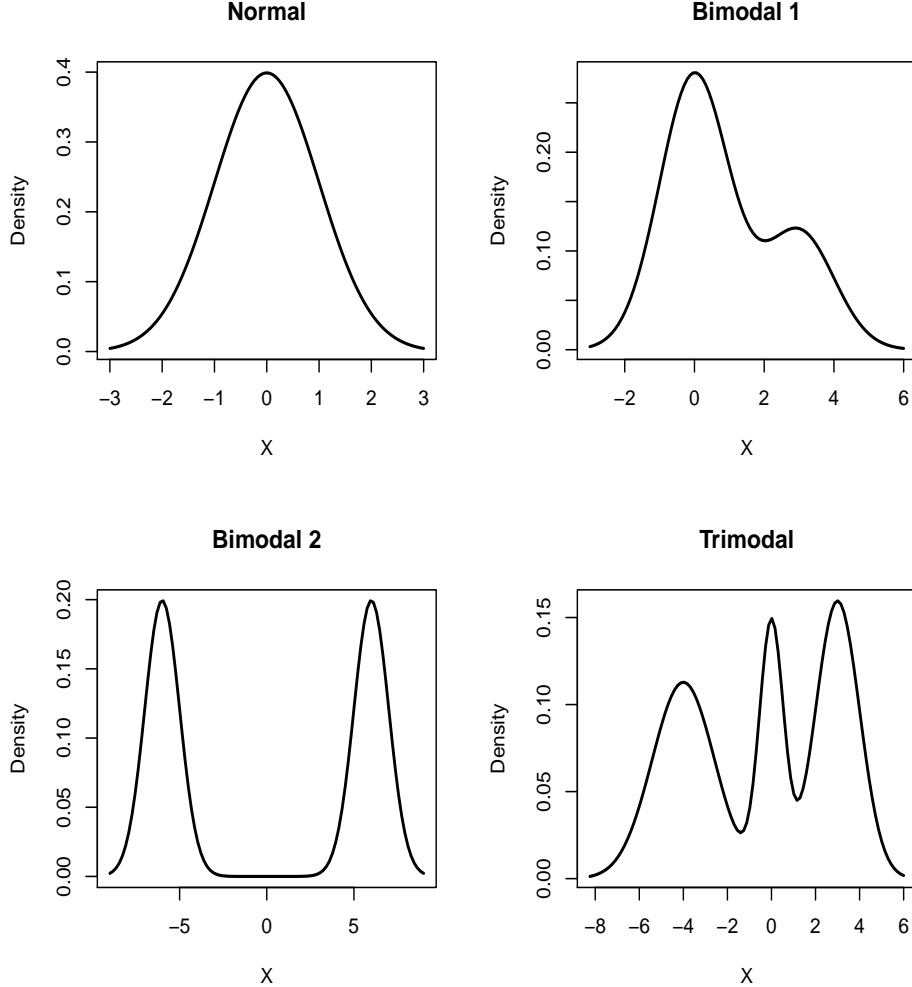


Figure 1: The four 1-dimensional Gaussian mixture densities we consider.

We now compare the MISE for the densities in Figure 1 using the three approaches for selecting the bandwidth parameter. Recall that the approaches are: 1) optimize the bandwidth for estimating f_Y , 2) optimize the bandwidth for estimating f_X , and 3) set the bandwidth equal to 0. In Section 3 we showed that optimizing the bandwidth for f_Y (approach 1) and setting the bandwidth equal to 0 (approach 3) resulted in the same first order asymptotic performance for the MISE. In contrast, optimizing the bandwidth for f_X

Name	Parameters
Normal	$\phi_1(x)$
Bimodal 1	$.7\phi_1(x) + .3\phi_1(x - 3)$
Bimodal 2	$.5\phi_1(x + 6) + .5\phi_1(x - 6)$
Trimodal	$.4\phi_2(x + 4) + .2\phi_{.3}(x) + .4\phi_1(x - 3)$

Table 1: Parameters for the four densities plotted in Figure 1 where ϕ_Σ is the normal, mean 0 density with covariance Σ .

(approach 2) resulted in slower, nonparametric convergence rates.

Let h_Y be the optimal bandwidth for estimating f_Y and h_X be the optimal bandwidth for estimating f_X (h_Y and h_X are sequences implicitly indexed by the sample size n). We now compare $MISE(h_Y)$, $MISE(h_X)$, and $MISE(0)$ at finite n for the four densities in Figure 1 and a variety of error variances σ_ϵ^2 . Clearly $MISE(h_Y) \leq MISE(h_X)$ and $MISE(h_Y) \leq MISE(0)$ since h_Y is the minimizer of the MISE. We seek to understand the level of reduction in MISE one can achieve by using h_Y , the parameter settings where these reductions occur, and how $MISE(h_X)$ compares to $MISE(0)$. We note that h_X and h_Y are exact minimizers for the MISE of f_Y and f_X , not asymptotic approximations.

In Table 2 we present $\left(\frac{MISE(0)}{MISE(h_Y)}, \frac{MISE(h_X)}{MISE(h_Y)}\right)$ for four densities and five error variances (the error is normal, mean 0) for $n = 50$. We note some general trends. As σ_ϵ^2 decreases, $MISE(h_X)/MISE(h_Y)$ decreases. With small σ_ϵ^2 , f_Y is close to f_X and thus h_X and h_Y are close. In contrast, as σ_ϵ^2 decreases, $MISE(0)/MISE(h_Y)$ increases. With no smoothing and small σ_ϵ^2 , approach 3 undersmooths the density estimate.

σ_ϵ^2	Normal	Bimodal 1	Bimodal 2	Trimodal
2	(1.02,1.18)	(1.08,1.01)	(1.03,1.02)	(1.18,1.05)
1	(1.05,1.17)	(1.15,1.01)	(1.07,1.03)	(1.24,1.04)
0.5	(1.13,1.11)	(1.26,1.01)	(1.16,1.03)	(1.30,1.01)
0.25	(1.32,1.05)	(1.50,1.00)	(1.37,1.01)	(1.46,1.00)
0.125	(1.70,1.02)	(1.92,1.00)	(1.76,1.01)	(1.77,1.00)

Table 2: Each entry is $\left(\frac{MISE(0)}{MISE(h_Y)}, \frac{MISE(h_X)}{MISE(h_Y)}\right)$ for $n = 50$. These ratios are always greater than 1 because h_Y is the minimizer of the MISE. As expected, $MISE(0)$ performs well when σ_ϵ^2 (the error variance) is large but poorly when σ_ϵ^2 is small. $MISE(h_X)$ performs well when σ_ϵ^2 is small but poorly when σ_ϵ^2 is large.

For the densities and error variances considered, using h_X (approach 2) is generally better than no

smoothing (approach 3). Only for the normal distribution with $\sigma_\epsilon^2 = 2$ or 1 does no smoothing outperform smoothing with h_X . This is surprising given the asymptotic results showed that the convergence rate using h_X is slower than the rate using no smoothing. For the densities and error distributions considered, a sample size of 50 is not large enough for these asymptotics to take hold. An important caveat to this conclusion is that the no smoothing estimator is simpler than h_X because it has no smoothing parameters.

σ_ϵ^2	Normal	Bimodal 1	Bimodal 2	Trimodal
2	(1.01,1.24)	(1.04,1.03)	(1.02,1.04)	(1.09,1.02)
1	(1.03,1.24)	(1.08,1.03)	(1.04,1.06)	(1.12,1.01)
0.5	(1.07,1.18)	(1.15,1.03)	(1.09,1.06)	(1.16,1.00)
0.25	(1.19,1.09)	(1.31,1.02)	(1.24,1.03)	(1.27,1.00)
0.125	(1.46,1.04)	(1.62,1.01)	(1.53,1.01)	(1.50,1.00)

Table 3: The entries here are the same as Table 2 but for $n = 100$. This larger n generally improves performance for $MISE(0)$ and worsens the performance of $MISE(h_X)$ (relative to $MISE(h_Y)$). This is predicted by our asymptotic theory, since as $n \rightarrow \infty$, $\frac{MISE(0)}{MISE(f_Y)} \rightarrow 1$ while $\frac{MISE(h_X)}{MISE(h_Y)} \rightarrow \infty$. However at $n = 100$, using h_X still generally outperforms no smoothing.

In Table 3 we plot the same quantities as Table 2 but for $n = 100$. The same general trends apply here as with the $n = 50$ case: As σ_ϵ^2 decreases $MISE(h_X)$ decreases while $MISE(0)$ increases (relative to $MISE(h_Y)$). Note that the performance of no smoothing relative to h_X is generally better for $n = 100$ than $n = 50$. For example, with Bimodal 2 and $\sigma_\epsilon^2 = 2$, $MISE(h_X) < MISE(0)$ for $n = 50$, but $MISE(h_X) > MISE(0)$ for $n = 100$. In fact, for every case considered $MISE(0)/MISE(h_Y)$ is lower for $n = 100$ than $n = 50$. With the exception of Trimodal, $MISE(h_X)/MISE(h_Y)$ is always greater for $n = 100$ than $n = 50$.

The asymptotic results in Section 3.1 predict this behavior. As $n \rightarrow \infty$, $\frac{MISE(0)}{MISE(f_Y)} \rightarrow 1$ while $\frac{MISE(h_X)}{MISE(h_Y)} \rightarrow \infty$. So for large enough n , $MISE(0) < MISE(h_X)$. However, for most densities and error variances considered here, $n = 100$ is not large enough for these asymptotics to take hold. These results suggest that at sample sizes of potential interest, using no smoothing can undersmooth the density estimate. This effect appears most significant when the error density is concentrated near 0.

4.2 Relative Error in Three Dimensions

We now explore relative error rates for 3-dimensional densities. The Gaussian mixture densities we study are defined in Table 4 using the notation

$$+ = \begin{bmatrix} 1 & 0.64 & 0 \\ 0.64 & 1 & 0.64 \\ 0 & 0.64 & 1 \end{bmatrix} \quad - = \begin{bmatrix} 1 & -0.64 & 0 \\ -0.64 & 1 & -0.64 \\ 0 & -0.64 & 1 \end{bmatrix} \quad (16)$$

for covariances matrices. The four densities in Table 4 are meant to be 3-dimensional generalizations of the 1-dimensional densities from Table 1. In particular Multi. Normal is a 3-dimensional normal, a direct generalization of the 1-dimensional normal. Multi 2-Comp 2 is a mixture of two well separated, identity covariance normals. This is a close analogue to Bimodal 2 from Table 1.

Name	Parameters
Multi. Normal	$\phi_I(x)$
Multi. 2-Comp 1	$.7\phi_+(x) + .3\phi_-(x - (1, 1, 1)^T)$
Multi. 2-Comp 2	$.5\phi_I(x - (6, 0, 0)^T) + .5\phi_I(x + (6, 0, 0)^T)$
Multi. 3-Comp	$.4\phi_+(x) + .2\phi_-(x - (1, 1, 1)^T) + .4\phi_-(x)$

Table 4: Parameters for the four 3-dimensional densities studied. ϕ_Σ is the normal, mean 0 density in three dimensions with covariance Σ . Here I is the 3×3 identity matrix and the $+$ and $-$ signs are covariance matrices defined in Equation 16.

We consider 5 error densities for ϵ . Each error density is normal, mean 0, with all diagonal elements equal and 0 for all covariances. The normality of ϵ is required by Theorem 3. The other choices were made to keep these simulations a reasonable size. For diagonal terms of the covariance, we consider the same values as for the 1-dimensional case: 2, 1, 0.5, 0.25, and 0.125.

Tables 5 and 6 present the ratios $\left(\frac{MISE(0)}{MISE(h_Y)}, \frac{MISE(h_X)}{MISE(h_Y)} \right)$ for all 20 f_X, f_ϵ pairs for $n = 100$ and $n = 500$ respectively. The first column in each table, σ_ϵ^2 , refers to the diagonal elements of the covariance matrix. The $n = 100$ case, Table 5, allows for direct comparison with the 1-dimensional case in Table 3. The $n = 500$ case, Table 6, provides results for what is perhaps a more realistic sample size when attempting to estimate a 3-dimensional density non-parametrically.

We discuss the $n = 100$ results, Table 5. In general h_Y performs better relative to no smoothing and h_X smoothing in three dimensions than in one dimension. In particular, all ratios are larger for Multi. Normal and Multi. 2-Comp 2 in Table 5 than Normal and Bimodal 2 in Table 3. As before, with small error

variance, approach 3 undersmooths the density estimate. As in the 1-dimensional case, h_X oversmooths the density estimates when the error variance is large. This effect appears worse in three dimensions than in one dimension. For example, for the standard normal with $\sigma_\epsilon^2 = 2$ and $n = 100$, $MISE(h_X)/MISE(h_Y)$ is 1.24 in one dimension (Table 3) and 1.76 in three dimensions (Table 5).

σ_ϵ^2	Multi Normal	Multi 2-Comp 1	Multi 2-Comp 2	Multi 3-Comp
2	(1.02,1.76)	(1.02,1.13)	(1.04,1.28)	(1.02,1.20)
1	(1.07,1.63)	(1.06,1.12)	(1.12,1.28)	(1.07,1.15)
0.5	(1.24,1.35)	(1.16,1.08)	(1.39,1.17)	(1.21,1.07)
0.25	(1.80,1.14)	(1.40,1.05)	(2.18,1.07)	(1.55,1.02)
0.125	(3.38,1.05)	(2.00,1.02)	(4.34,1.02)	(2.32,1.01)

Table 5: Three dimensional finite sample results for $n = 100$. Generally, h_X and no smoothing perform worse relative to h_Y here than for $n = 100$ in one dimension (see Table 3).

We now discuss the results for $n = 500$, Table 6. Note that with the larger sample size, all of the $MISE(0)/MISE(h_Y)$ ratios have decreased while all of the $MISE(h_X)/MISE(h_Y)$ ratios have increased relative to the $n = 100$ case in Table 5. The asymptotic results from Section 3 predict this behavior. As $n \rightarrow \infty$, $MISE(0)/MISE(h_Y) \rightarrow 1$ while $MISE(h_X)/MISE(h_Y) \rightarrow \infty$. As expected, for all densities $MISE(0)/MISE(h_Y)$ is increasing as σ_ϵ^2 decreases.

σ_ϵ^2	Multi Normal	Multi 2-Comp 1	Multi 2-Comp 2	Multi 3-Comp
2	(1.00,2.66)	(1.00,1.27)	(1.01,1.72)	(1.01,1.37)
1	(1.01,2.54)	(1.01,1.30)	(1.03,1.82)	(1.02,1.34)
0.5	(1.06,2.00)	(1.03,1.29)	(1.10,1.57)	(1.05,1.25)
0.25	(1.25,1.45)	(1.10,1.23)	(1.41,1.26)	(1.14,1.16)
0.125	(1.94,1.16)	(1.30,1.14)	(2.37,1.09)	(1.41,1.09)

Table 6: Three dimensional finite sample results for $n = 500$. $MISE(h_X)/MISE(h_Y)$ is larger and $MISE(0)/MISE(h_Y)$ is smaller here relative to Table 5 where the sample size was 100.

The three dimensional finite sample results reinforce the conclusion from the one dimensional results that not smoothing when the error variance is small produces undersmoothed estimates with large MISE relative to using the optimal smoothing parameter. Additionally, in three dimensions, using h_X oversmooths the density estimates in many cases where the error variance is large.

The 3-dimensional simulations here were limited to a very narrow set of error distributions. In particular

the error variance was equal in all directions. Another interesting setting to consider is when the error variance is very small along certain directions and sizable along other directions. The limiting case of this setting has no error along certain directions.

When there is no error in certain direction but error in other directions, one can obtain convergence rates that depend on the dimension of the space on which there is error (Long [2013], Theorem 2.3). Specifically, suppose one estimates a p dimensional density f_Y and ϵ is 0 with probability 1 on a p_0 dimensional subspace of \mathbb{R}^p . Suppose ϵ has a density on the other p_1 dimensions ($p = p_0 + p_1$). Then for second order kernels, under some regularity conditions, the optimal smoothing using a scalar bandwidth is of order $n^{-1/(4+p_1)}$ and results in an MISE of order $n^{-4/(4+p_1)}$ (see Equation 2.27 in Long [2013]). Note that this rate is between the error free rate of $n^{-4/(4+p)}$ and the error in all directions case where the MISE is of order n^{-1} .

4.3 Qualitative Impacts of Smoothing Parameters

We now visualize some of the results in Table 2 by plotting pointwise quantiles for density estimates for different choices of smoothing parameters. In order to obtain an understanding of the impact of using h_X (approach 2) or no smoothing (approach 3), we examine the 1-dimensional cases where these methods perform worst relative to using h_Y (approach 1). This shows some of the qualitative impacts of suboptimal smoothing on the density estimate.

We first study $\sigma_\epsilon^2 = 2$, Normal. Here h_X performed worst (relative to h_Y) out of all the densities and error variances considered (see Table 2). We generate 100 samples of size 50 from Normal. Using these 100 samples, we construct 100 density estimates using h_Y and h_X . In Figure 2 a) we plot the .1 and .9 pointwise quantiles for these density estimates (orange-dashed for h_Y and blue-dotted for h_X) along with the true underlying density f_Y (i.e. the Normal density convolved with ϕ_2) in black-solid. The quantiles for h_X have a lower peak and heavier tails than the quantiles for the h_Y density estimates. Using h_X oversmooths the density estimates ($h_X = 0.52$ and $h_Y = 0.26$).

In Figures 2 b) and c) we plot 10 density estimates using h_Y and h_X respectively. We see that the individual density estimates using h_X are negatively biased near $Y = 0$ and positively biased for large $|Y|$. Since all the density estimates are unimodal, with a mode near 0 and approximately normal, the qualitative conclusions that one is likely to draw from these density estimates are likely to be similar regardless of whether one is using h_X or h_Y .

We now study the Bimodal 1 density case with $\sigma_\epsilon^2 = .125$ and $n = 50$. Here $MISE(0)/MISE(h_Y)$ was

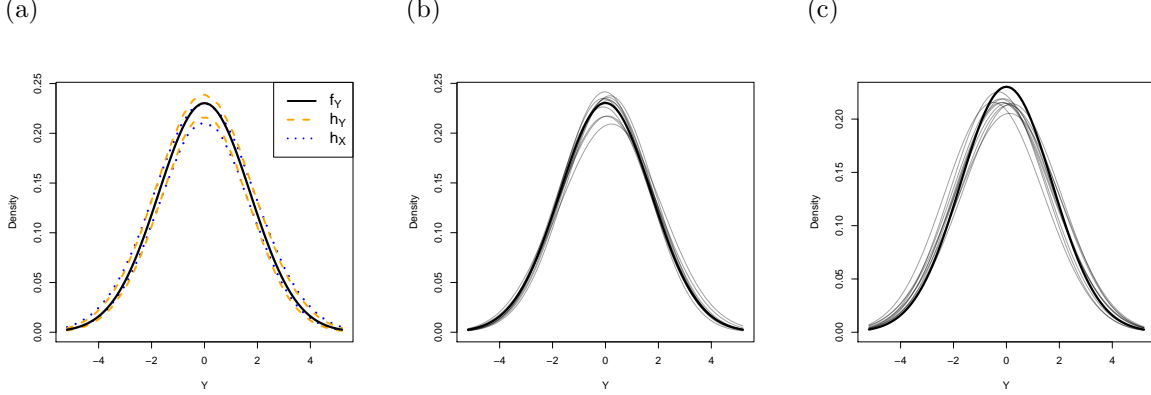


Figure 2: Comparison of using h_Y (optimal smoothing for f_Y) to h_X (optimal smoothing for f_X) for the Normal density with $\sigma_\epsilon^2 = 2$. In a) we plot f_Y and the .9 and .1 quantiles for density estimates using h_Y (orange-dash) and h_X (blue-dot). h_X oversmooths the estimate, so the peak at $Y = 0$ is biased low while the tails are biased high. In b) and c) we plot 10 density estimates using h_Y and h_X respectively. The qualitative conclusions that one is likely to draw from the density estimates are similar, regardless of whether h_X or h_Y is used.

highest out of all conditions tested in Table 2. Following the procedure for generating Figure 2, we generate 100 samples of size 50 from Bimodal 1. Using these 100 samples, we construct 100 density estimates using h_Y and no smoothing. In Figure 3 a) we plot the .1 and .9 pointwise quantiles for these density estimates (orange-dashed for h_Y and blue-dotted for no smoothing). We plot the true underlying density, f_Y (i.e. the Bimodal 1 density convolved with $\phi_{.125}$), in black-solid.

No smoothing greatly overestimates the height of the mode at $Y = 0$. The quantiles for the h_Y density estimates are nearly contained within the quantiles for no smoothing across all values of Y . In Figure 3 b) and c) we plot 10 density estimates using h_Y and no smoothing respectively. The density estimates using h_Y (in b)) typically identify two modes close to the correct Y values. In contrast the density estimates using no smoothing appear very undersmoothed (in c)). Several estimates have three or more modes and the mode heights are often far from the true value. In this case, not smoothing could have a significant impact on qualitative conclusions drawn from the density estimate. The results from Figures 2 and 3 suggest that the qualitative impacts of not smoothing may be worse than smoothing using h_X .

4.4 Convergence of Bandwidth to Asymptotic Approximation

We now study the rate of convergence of the asymptotically optimal bandwidth for estimating f_Y , h_Y^* (see Equation 13) to the exact optimal bandwidth h_Y . Fast convergence rates suggest that plug-in estimators

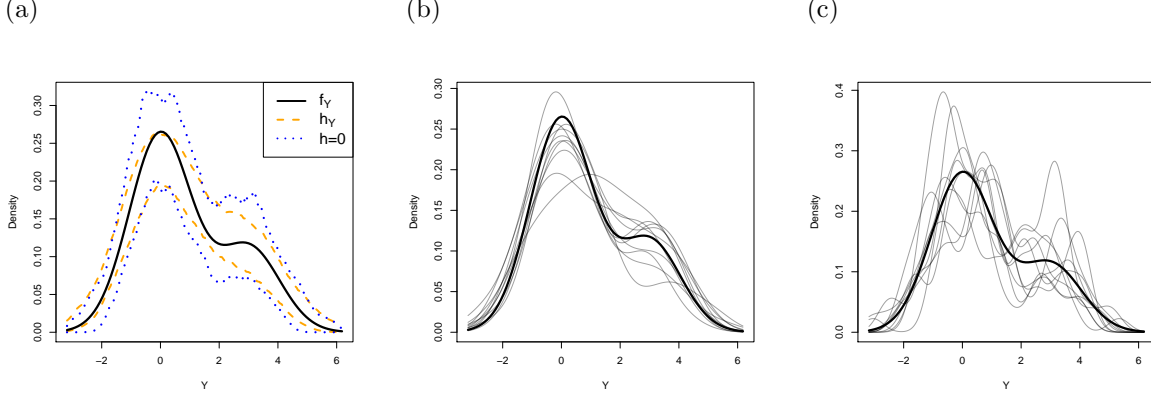


Figure 3: Comparison of using h_Y (optimal smoothing for f_Y) to no smoothing for the Bimodal 1 density with $\sigma_\epsilon^2 = .125$. In a) we plot f_Y (black-solid) and the .9 and .1 quantiles for density estimates using h_Y (orange-dash) and no smoothing (blue-dot). The quantiles for no smoothing are wider than for h_Y for most values of Y . In b) and c) we plot 10 density estimates using h_Y and no smoothing respectively. The density estimates using no smoothing often have 3 modes. These modes are often not close to the true Y value modes.

could be effective for estimating h_Y . We pay particular attention to the relationship between convergence rate of h_Y^* and the error variance σ_ϵ^2 .

For the four densities in Figure 1 using error variances $\sigma_\epsilon^2 = 2, 1, .5, .25, .125$ we compute the ratio between the exact optimal bandwidth (h_Y) and the asymptotically optimal bandwidth (h_Y^*). The exact optimal bandwidth is determined by finding the h which minimizes Equation 15. The asymptotically optimal bandwidth is computed using Equation 13. We plot these ratios as a function of n for the Normal and Trimodal densities in Figure 4 a) and b) respectively.

For the Normal density, the larger σ_ϵ^2 , the faster the convergence of the asymptotically optimal bandwidth to the actual optimal bandwidth. For example with $\sigma_\epsilon^2 = 2, 1, .5$ and $n = 100$, the exact optimal bandwidth is within 10% of the asymptotic expression. This suggests that for a normal density, with moderate n and error variance not too small, plug-in estimators for the asymptotic bandwidth may provide a good approximation to the bandwidth which minimizes the exact MISE. The plots for Bimodal 1 and Bimodal 2 (not shown) closely resemble the Normal density.

For the Trimodal density, the relationship between σ_ϵ^2 and the convergence rate of the asymptotically optimal bandwidth (h_Y^*) to exact optimal bandwidth (h_Y) is more complicated than for the Normal, Bimodal 1, and Bimodal 2 densities. Broadly, the asymptotics for $\sigma_\epsilon^2 = .25, .125$ take hold at larger n than for $\sigma_\epsilon^2 = 2, 1, .5$. Unlike for the normal case, the asymptotically optimal bandwidth is not always greater than

the exact optimal bandwidth.

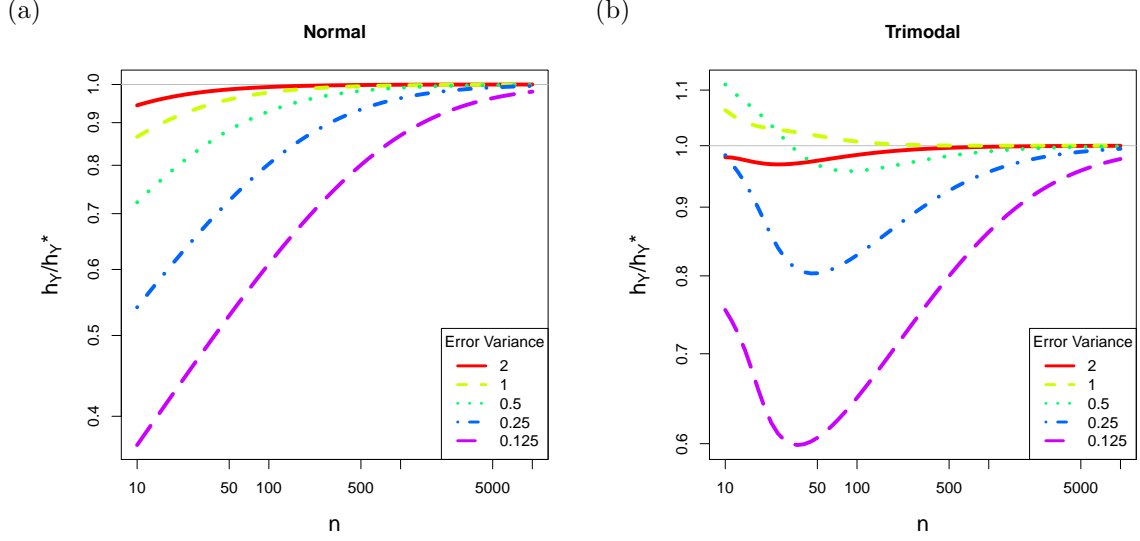


Figure 4: Ratio of exact optimal bandwidth to asymptotically optimal bandwidth (h_Y/h_Y^*) as a function of n for the Normal (a) and Trimodal (b) densities. The convergence of this ratio to 1 varies with σ_ϵ^2 (the error variance). For the Normal density the relationship between σ_ϵ^2 and the convergence is fairly simple, while for the Trimodal, the behavior is more complex.

Certain aspects of the convergence rate behavior in Figure 4 may be explained by considering asymptotics in σ_ϵ^2 . At constant n , as $\sigma_\epsilon^2 \rightarrow 0$, $h_Y/h_Y^* \rightarrow 0$. This can be seen by considering the limiting values (in σ_ϵ^2) of h_Y and h_Y^* . Note that as $\sigma_\epsilon^2 \rightarrow 0$, $f_Y \rightarrow f_X$. Therefore $h_Y \rightarrow h_X$, a positive constant. In contrast, as $\sigma_\epsilon^2 \rightarrow 0$, $h_Y^* = \sqrt{\frac{2 \int (\omega^T \Sigma_K \omega) d\nu(\omega)}{n \int (\omega^T \Sigma_K \omega)^2 d\mu(\omega)}} \rightarrow \infty$. This behavior is seen in Figure 4 a) where at fixed n , the smaller σ_ϵ^2 , the smaller h_Y/h_Y^* . For $n > 1000$, this relationship between σ_ϵ^2 and h_Y/h_Y^* is true for Trimodal (Figure 4 b)) as well. Consideration of asymptotic regimes in which $n \rightarrow \infty$ and $\sigma_\epsilon^2 \rightarrow 0$ together at some rate may help explain the behavior in Figure 4 better.

In cases where the asymptotically optimal bandwidth is far from the exact optimal bandwidth, estimators for the asymptotically optimal bandwidth (such as plug-in or “rule-of-thumb” estimators) may have poor performance in terms of minimizing risk. The simulation results in Figure 4 suggest that when ϵ is concentrated around 0, h_Y^* is far from h_Y . Thus plug-in estimators, which attempt to estimate h_Y^* , may be suboptimal in terms of minimizing MISE. In Section 5 we observe this behavior with a “rule-of-thumb” estimator for h_Y when ϵ has small variance.

5 Estimator for h_Y and Real Data Example

5.1 Rule-of-Thumb Estimator for h_Y

Jones et al. [1996] define “rule-of-thumb” bandwidth selection procedures as any method which replaces unknown quantities in the asymptotically optimal bandwidth with estimated values based on a parametric family for the unknown density. We now propose a rule-of-thumb estimation method for h_Y . Recall from Equation (13) that the asymptotically optimal bandwidth is

$$h_Y^* = \sqrt{\frac{2 \int (\omega^T \Sigma_K \omega) d\nu(\omega)}{n \int (\omega^T \Sigma_K \omega)^2 d\mu(\omega)}} = \sqrt{\frac{2 \int (\omega^T \Sigma_K \omega) |\hat{f}_\epsilon(\omega)|^2 (1 - |\hat{f}_X(\omega)|^2) d\omega}{n \int (\omega^T \Sigma_K \omega)^2 |\hat{f}_X(\omega)|^2 |\hat{f}_\epsilon(\omega)|^2 d\omega}}. \quad (17)$$

We specialize to one dimension, so Σ_K is a scalar and can be set to 1 without loss of generality. h_Y^* depends on $|\hat{f}_X(\omega)|^2$, which is unknown. We replace this quantity by assuming (solely for the purposes of bandwidth estimation) that f_X is mean 0 normal. In this case, $|\hat{f}_X(\omega)|^2 = e^{-\sigma_X^2 \omega^2}$ where σ_X^2 is the variance of f_X . σ_X^2 is estimated with $\tilde{\sigma}_X^2$, the variance of the observations X_1, \dots, X_n . Thus our rule-of-thumb bandwidth estimator for Berkson kernel density estimation is

$$\tilde{h}_Y = \sqrt{\frac{2 \int \omega^2 |\hat{f}_\epsilon(\omega)|^2 (1 - e^{-\tilde{\sigma}_X^2 \omega^2}) d\omega}{n \int \omega^4 e^{-\tilde{\sigma}_X^2 \omega^2} |\hat{f}_\epsilon(\omega)|^2 d\omega}}. \quad (18)$$

For the case where ϵ is mean 0 normal with variance σ_ϵ^2 , $|\hat{f}_\epsilon(\omega)|^2 = e^{-\sigma_\epsilon^2 \omega^2}$ and Equation 18 simplifies to

$$\tilde{h}_Y = \sqrt{\frac{4}{3n} \left[\frac{(\tilde{\sigma}_X^2 + \sigma_\epsilon^2)^{5/2}}{\sigma_\epsilon^2} - (\sigma_\epsilon^2 + \tilde{\sigma}_X^2) \right]}. \quad (19)$$

5.2 Real Data Example

We analyze data collected by Ferris Jr et al. [1979] concerning childhood exposure to NO_2 , a known cause of respiratory illness. The goal is to determine the density of exposure to NO_2 for children living in Watertown, Massachusetts. Ferris Jr et al. [1979] collected kitchen and bathroom concentrations of NO_2 for 231 homes in Watertown. In this study, direct personal exposure to NO_2 was not observed.

Using data collected in Portage, Wisconsin and the Netherlands, Tosteson et al. [1989] modeled log personal exposure to NO_2 (Y) as a linear function of log kitchen ($\ln(W_k)$) and log bathroom ($\ln(W_b)$) concentrations plus random error. Specifically (see Table 1 of Tosteson et al. [1989])

$$Y = 1.22 + 0.3 \ln(W_k) + 0.33 \ln(W_b) + \epsilon$$

where $\epsilon \sim N(0, .06)$, independent of W_k and W_b . Let $X = 1.22 + 0.3 \ln(W_k) + 0.33 \ln(W_b)$. By assuming the same error model holds in Watertown as in Portage and the Netherlands (i.e. assuming portability of

the error model), we can estimate log personal exposure density in Watertown as X plus independent noise where we have 231 observed values of X .

We estimate the density of Y using the three smoothing approaches described in earlier sections: smoothing to optimize estimation of f_Y , smoothing to optimize estimation of f_X and no smoothing. For no smoothing we simply convolve the observations with the error density. For smoothing to optimize estimation of f_X and f_Y we must select a kernel K and bandwidth estimation methods for h_X and h_Y . We use a Gaussian kernel. For estimating h_X we use “Silverman’s rule-of-thumb”, developed in Deheuvels [1977] and Silverman [1986],⁵

$$\tilde{h}_X = \frac{0.9 \min(\widetilde{IQR}_X/1.34, \tilde{\sigma}_X)}{n^{1/5}}.$$

Here \widetilde{IQR} and $\tilde{\sigma}_X$ are the estimated inter-quartile range and standard deviation of X . For estimating h_Y we use the rule-of-thumb estimator \tilde{h}_Y proposed in Equation (19).

In addition to studying the case $\epsilon \sim N(0, .06)$, we construct density estimates when ϵ is normal with variance 0.6 and 0.006 in order to study robustness of the smoothing methods to different levels of Berkson error. In Figure 5 we plot the three estimators for a) $\epsilon \sim N(0, 0.6)$, b) $\epsilon \sim N(0, 0.06)$, and c) $\epsilon \sim N(0, 0.006)$. In a) where $\sigma_\epsilon^2 = 0.6$ there is essentially no difference in the estimators. In b) where $\sigma_\epsilon^2 = 0.06$ no smoothing results in a somewhat higher mode around $y = 3$ than smoothing to optimize estimation of f_X or f_Y . It appears unlikely that the choice of smoothing would affect qualitative conclusions in this case. In c) where $\sigma_\epsilon^2 = 0.006$ no smoothing severely under-regularizes the density estimate. In particular the estimate has four modes. \tilde{h}_Y oversmooths the estimate of f_Y . This is likely due to the fact that for fixed n as $\sigma_\epsilon \rightarrow 0$, $h_Y^* \rightarrow \infty$ while $h_Y \rightarrow h_X$ (see Subsection 4.4). Since \tilde{h}_Y is an estimate of h_Y^* , it oversmooths the density estimate in this case. Overall, with $\epsilon \sim N(0, 0.006)$, \tilde{h}_X produces the best density estimate (blue dashed line).

6 Conclusions

In this work we compared different approaches to smoothing a density estimate subject to Berkson error. No smoothing (approach 3) achieved suboptimal asymptotic (at second order) and finite sample MISE. This was especially evident when the error term ϵ was concentrated near 0. Smoothing to optimize estimation of f_X resulted in suboptimal asymptotic (at first order) MISE rates. At finite samples, h_X oversmoothed density

⁵This is the default bandwidth selection for the `density` function in the software package `R`, Version 3.01.

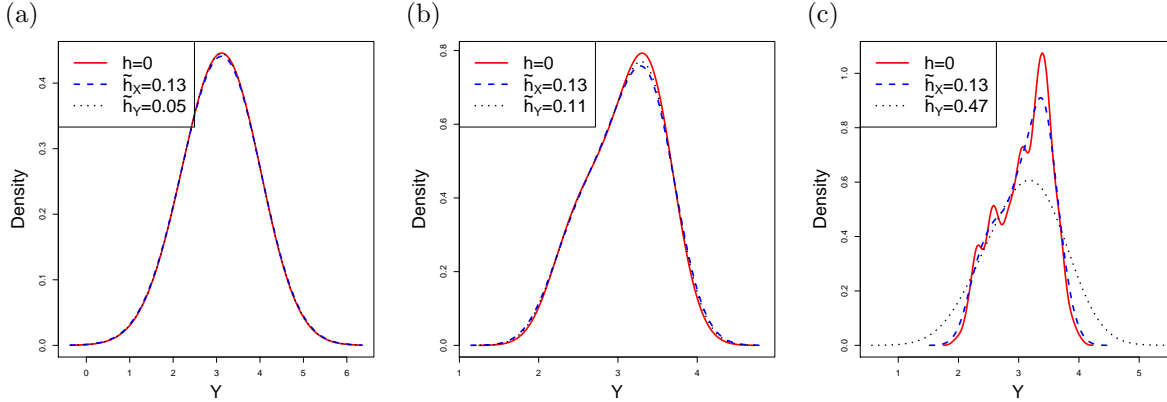


Figure 5: Density estimates of log NO₂ exposure for children living in Watertown using three different error variances. With error variances $\sigma_\epsilon^2 = 0.6$ or 0.06 , plots a) and b) respectively, all three smoothing methods produce similar density estimates. In c), where $\sigma_\epsilon^2 = 0.006$, no smoothing under-regularizes the density estimate.

estimates, particularly when the error variance was large. These effects grew worse in higher dimensions.

These results support using a bandwidth specifically chosen for estimation of f_Y . More work is needed to develop estimators of h_Y . The asymptotically optimal bandwidth h_Y^* derived in Equation (13) suggest one form for rule-of-thumb and plug-in type estimators. The simulations in Subsection 4.4 suggest that when the error is concentrated around 0, using a bandwidth which estimates h_Y^* (the asymptotic approximation to h_Y) may oversmooth the density estimate. The rule-of-thumb estimator for h_Y we developed in Equation (19) displayed this behavior. Estimators for h_Y based on cross-validation or the bootstrap may perform better under a wider range of possible error distributions and should be developed. The development of bandwidth estimators for standard kernel density estimation (see Jones et al. [1996] for a review) suggest forms for such procedures.

Acknowledgments

Support from NSF grant 0941742 (Cyber-Enabled Discovery and Innovation), NSF grant DMS-0847647 (CAREER), and a fellowship from Citadel LLC are gratefully acknowledged. We thank Raymond Carroll for providing helpful comments on the manuscript and Len Stefanski for supplying the data set on children's NO₂ exposure.

References

J. Berkson. Are there two regressions? *Journal of the American Statistical Association*, 45(250):164–180, 1950. ISSN 0162-1459.

- J. Bovy, J. F. Hennawi, D. W. Hogg, A. D. Myers, J. A. Kirkpatrick, D. J. Schlegel, N. P. Ross, E. S. Sheldon, I. D. McGreer, D. P. Schneider, et al. Think outside the color box: Probabilistic target selection and the SDSS-XDQSO Quasar targeting catalog. *The Astrophysical Journal*, 729(2):141, 2011.
- R. Carroll, D. Ruppert, L. Stefanski, and C. M. Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC Press, 2006. ISBN 1584886331.
- R. Carroll, A. Delaigle, and P. Hall. Nonparametric prediction in measurement error models. *Journal of the American Statistical Association*, 104(487):993–1003, 2009. ISSN 0162-1459.
- P. Deheuvels. Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée*, 25(3):5–42, 1977.
- A. Delaigle. Nonparametric density estimation from data with a mixture of berkson and classical errors. *Canadian Journal of Statistics*, 35(1):89–104, 2007.
- A. Delaigle. An alternative view of the deconvolution problem. *Statistica Sinica*, 18(3):1025–1045, 2008.
- R. Durrett. *Probability : theory and examples*. Duxbury advanced series. Brooks/Cole, Belmont, USA, 2005. ISBN 0-534-42441-4.
- C. H. Edwards Jr. *Advanced calculus of several variables*. Dover Publications, 1973.
- B. Ferris Jr, F. Speizer, J. Spengler, D. Dockery, Y. Bishop, M. Wolfson, C. Humble, et al. Effects of sulfur oxides and respirable particles on human health. methodology and demography of populations in study. *The American review of respiratory disease*, 120(4):767, 1979.
- H. Henderson and S. Searle. Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *Canadian Journal of Statistics*, 7(1):65–81, 1979.
- M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.
- J. P. Long. *Prediction Methods for Astronomical Data Observed with Measurement Error*. PhD thesis, University of California, Berkeley, 2013.
- J. P. Long, N. El Karoui, J. A. Rice, J. W. Richards, and J. S. Bloom. Optimizing automated classification of variable stars in new synoptic surveys. *Publications of the Astronomical Society of the Pacific*, 124(913): 280–295, 2012.
- J. S. Marron and M. P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736, 1992.
- B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.
- B. C. Sutradhar. On the characteristic function of multivariate student t-distribution. *Canadian Journal of Statistics*, 14(4):329–337, 1986.
- T. D. Tosteson, L. A. Stefanski, and D. W. Schafer. A measurement-error model for binary and ordinal regression. *Statistics in Medicine*, 8(9):1139–1147, 1989.
- A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009. ISBN 1441927093.
- N. G. Ushakov. *Selected topics in characteristic functions*. De Gruyter Mouton, 1999.

- M. Wand and M. Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, pages 520–528, 1993.
- M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60. Chapman & Hall/CRC, 1995.

Supplementary Material to *Kernel Density Estimation with Berkson Error*

James P. Long

Department of Statistics, Texas A&M University
3143 TAMU, College Station, TX 77843-3143
jlong@stat.tamu.edu

Nouredine El Karoui

Department of Statistics, University of California, Berkeley
367 Evans Hall # 3860, Berkeley, CA 94720-3860
nkaroui@stat.berkeley.edu

John A. Rice

Department of Statistics, University of California, Berkeley
367 Evans Hall # 3860, Berkeley, CA 94720-3860
rice@stat.berkeley.edu

S.1 Proofs of the Theorems

S.1.1 Proof of Theorem 1

We must show

$$(2\pi)^p \text{MISE}(H) = \int |1 - \widehat{K}(H\omega)|^2 d\mu(\omega) + \frac{1}{n} \int |\widehat{K}(H\omega)|^2 d\nu(\omega)$$

where

$$\begin{aligned} d\mu(\omega) &= |\widehat{f}_\epsilon(\omega)|^2 |\widehat{f}_X(\omega)|^2 d\omega, \\ d\nu(\omega) &= |\widehat{f}_\epsilon(\omega)|^2 (1 - |\widehat{f}_X(\omega)|^2) d\omega. \end{aligned}$$

Substituting for $d\mu(\omega)$ and $d\nu(\omega)$, it suffices to show that

$$(2\pi)^p \text{MISE}(H) = \int |\widehat{f}_\epsilon(\omega)|^2 \left(|1 - \widehat{K}(H\omega)|^2 |\widehat{f}_X(\omega)|^2 + \frac{1}{n} |\widehat{K}(H\omega)|^2 (1 - |\widehat{f}_X(\omega)|^2) \right) d\omega. \quad (\text{S.1})$$

$\widetilde{f}_{Y,H}, \widehat{f}_Y \in L_1$ by assumption. They are also in L_2 because they are characteristic functions and thus bounded. Under these conditions, the Plancherel theorem (see Theorem 1.8.8 on page 57 in Ushakov [1999]) states

$$\int (f_Y(y) - \widetilde{f}_{Y,H}(y))^2 dy = \frac{1}{(2\pi)^p} \int |\widehat{f}_Y(\omega) - \widetilde{\widehat{f}}_{Y,H}(\omega)|^2 d\omega. \quad (\text{S.2})$$

Let \mathcal{P}_n be the product measure on (X_1, \dots, X_n) . Using the definition of $\text{MISE}(H)$, Equation (S.2), and the facts $\widehat{f}_Y(\omega) = \widehat{f}_X(\omega)\widehat{f}_\epsilon(\omega)$ and $\widetilde{f}_{Y,H}(\omega) = \widehat{K}(H\omega)\widehat{f}_\epsilon(\omega)\widetilde{f}_X(\omega)$, we have

$$\begin{aligned} \text{MISE}(H) &= \mathbb{E}_{\mathcal{P}_n} \int \left(f_Y(y) - \widetilde{f}_{Y,H}(y) \right)^2 dy \\ &= \frac{1}{(2\pi)^p} \mathbb{E}_{\mathcal{P}_n} \int |\widehat{f}_Y(\omega) - \widetilde{f}_{Y,H}(\omega)|^2 d\omega \\ &= \frac{1}{(2\pi)^p} \mathbb{E}_{\mathcal{P}_n} \int |\widehat{K}(H\omega)\widehat{f}_\epsilon(\omega)\widetilde{f}_X(\omega) - \widehat{f}_X(\omega)\widehat{f}_\epsilon(\omega)|^2 d\omega \\ &= \frac{1}{(2\pi)^p} \mathbb{E}_{\mathcal{P}_n} \int |\widehat{f}_\epsilon(\omega)|^2 |\widetilde{f}_X(\omega)\widehat{K}(H\omega) - \widehat{f}_X(\omega)|^2 d\omega. \end{aligned}$$

Note that the integrand is a non-negative function, so we move the expectation inside the integral using Fubini's Theorem. We have

$$(2\pi)^p \text{MISE}(H) = \int |\widehat{f}_\epsilon(\omega)|^2 \mathbb{E}_{\mathcal{P}_n} |\widetilde{f}_X(\omega)\widehat{K}(H\omega) - \widehat{f}_X(\omega)|^2 d\omega.$$

Noting that it is sufficient to show Equation (S.1) holds, all that is left is to show is

$$\mathbb{E}_{\mathcal{P}_n} |\widetilde{f}_X(\omega)\widehat{K}(H\omega) - \widehat{f}_X(\omega)|^2 = |1 - \widehat{K}(H\omega)|^2 |\widehat{f}_X(\omega)|^2 + \frac{1}{n} |\widehat{K}(H\omega)|^2 (1 - |\widehat{f}_X(\omega)|^2).$$

This identity is shown in the proof of Theorem 1.4 on page 22 in Tsybakov [2009]. \square

S.1.2 Proof of Theorem 2

Recall that we are working under Assumptions A and B. This proof is divided into three parts. In **Part 1** we show $\widehat{f}_Y, \widetilde{f}_{Y,H} \in L_1$, which satisfies the conditions for Theorem 1 and implies

$$(2\pi)^p \text{MISE}(H) = \int |1 - \widehat{K}(H\omega)|^2 d\mu(\omega) + \frac{1}{n} \int |\widehat{K}(H\omega)|^2 d\nu(\omega). \quad (\text{S.3})$$

In **Part 2** we expand the first term of the right hand side of Equation (S.3) to show

$$\int |1 - \widehat{K}(H\omega)|^2 d\mu(\omega) = \left(\frac{1}{4} \int (\omega^T H^T \Sigma_K H \omega)^2 d\mu(\omega) \right) (1 + O(\|H\|_\infty^2)). \quad (\text{S.4})$$

In **Part 3** we expand the second term of the right hand side of Equation (S.3) to show

$$\frac{1}{n} \int |\widehat{K}(H\omega)|^2 d\nu(\omega) = \frac{1}{n} \int d\nu(\omega) - \left(\frac{1}{n} \int (\omega^T H^T \Sigma_K H \omega) d\nu(\omega) \right) (1 + O(\|H\|_\infty^2)). \quad (\text{S.5})$$

Summing Equations (S.4) and (S.5) we have the result

$$\begin{aligned} (2\pi)^p \text{MISE}(H) &= \frac{1}{n} \int d\nu(\omega) + \\ &+ \left(\frac{1}{4} \int (\omega^T H^T \Sigma_K H \omega)^2 d\mu(\omega) - \frac{1}{n} \int (\omega^T H^T \Sigma_K H \omega) d\nu(\omega) \right) (1 + O(\|H\|_\infty^2)). \end{aligned}$$

Part 1: $\widehat{f}_Y, \widetilde{f}_{Y,H} \in L_1$

Note that since the modulus of a characteristic function is bounded by 1

$$\begin{aligned} |\widehat{f}_Y(\omega)| &= |\widehat{f}_X(\omega)\widehat{f}_\epsilon(\omega)| \leq |\widehat{f}_\epsilon(\omega)|, \\ |\widetilde{f}_{Y,H}(\omega)| &= |\widehat{K}(H\omega)\widehat{f}_\epsilon(\omega)\widetilde{f}_X(\omega)| \leq |\widehat{f}_\epsilon(\omega)|. \end{aligned}$$

$\widehat{f}_\epsilon \in L_1$ by Assumption (10), implying $\widehat{f}_Y, \widetilde{\widehat{f}}_{Y,H} \in L_1$.

Part 2: Bias

By Lemma 1 on p.S.6 there exists R satisfying

$$|R(\omega)| \leq C\|\omega\|_\infty^4 \quad (\text{S.6})$$

such that

$$\widehat{K}(\omega) = 1 - \frac{\omega^T \Sigma_K \omega}{2} + R(\omega). \quad (\text{S.7})$$

Note that the kernel K is symmetric so \widehat{K} and R are real valued functions.

$$\begin{aligned} \int |1 - \widehat{K}(H\omega)|^2 d\mu(\omega) &= \int \left| \frac{\omega^T H^T \Sigma_K H \omega}{2} - R(H\omega) \right|^2 d\mu(\omega) \\ &= \frac{1}{4} \int (\omega^T H^T \Sigma_K H \omega)^2 d\mu(\omega) \\ &\quad - \int R(H\omega) (\omega^T H^T \Sigma_K H \omega) d\mu(\omega) \end{aligned} \quad (\text{S.8})$$

$$+ \int R(H\omega)^2 d\mu(\omega). \quad (\text{S.9})$$

We have split the integrals formally. We now show that Expressions (S.8) and (S.9) are $O(\|H\|_\infty^6)$ by bounding their integrands. Using the bound $R(\omega) \leq C\|\omega\|_\infty^4$ (Equation (S.6)), for some E we have

$$|R(H\omega) (\omega^T H^T \Sigma_K H \omega)| \leq C\|H\omega\|_\infty^4 \|\omega^T H^T \Sigma_K H \omega\|_\infty \leq E\|H\|_\infty^6 \|\omega\|_\infty^6,$$

$$|R(H\omega)^2| \leq C^2\|H\omega\|_\infty^8 \leq E\|H\|_\infty^8 \|\omega\|_\infty^8.$$

Using the definition of $d\mu(\omega)$ and the fact $\int \|\omega\|_\infty^8 |\widehat{f}_\epsilon(\omega)| d\omega < \infty$ (Assumption (9)) we have

$$\int \|\omega\|_\infty^8 d\mu(\omega) = \int \|\omega\|_\infty^8 |\widehat{f}_X(\omega)|^2 |\widehat{f}_\epsilon(\omega)|^2 d\omega \leq \int \|\omega\|_\infty^8 |\widehat{f}_\epsilon(\omega)|^2 d\omega < \infty.$$

So Expressions (S.8) and (S.9) are $O(\|H\|_\infty^6)$ and $O(\|H\|_\infty^8)$ respectively. Thus

$$\int |1 - \widehat{K}(H\omega)|^2 d\mu(\omega) = \left(\frac{1}{4} \int (\omega^T H^T \Sigma_K H \omega)^2 d\mu(\omega) \right) (1 + O(\|H\|_\infty^2)).$$

Part 3: Variance Using the expansion of \widehat{K} in Equation (S.7) we have

$$\frac{1}{n} \int |\widehat{K}(H\omega)|^2 d\nu(\omega) = \frac{1}{n} \int \left| 1 - \frac{\omega^T H^T \Sigma_K H \omega}{2} + R(H\omega) \right|^2 d\nu(\omega).$$

Expanding the right hand side we have

$$\frac{1}{n} \int \left| 1 - \frac{\omega^T H^T \Sigma_K H \omega}{2} + R(H\omega) \right|^2 d\nu(\omega) = \frac{1}{n} \left(\int d\nu(\omega) \right. \quad (\text{S.10})$$

$$- \int (\omega^T H^T \Sigma_K H \omega) d\nu(\omega) \quad (\text{S.11})$$

$$+ \frac{1}{4} \int (\omega^T H^T \Sigma_K H \omega)^2 d\nu(\omega) \quad (\text{S.12})$$

$$- \int R(H\omega) (\omega^T H^T \Sigma_K H \omega) d\nu(\omega) \quad (\text{S.13})$$

$$+ 2 \int R(H\omega) d\nu(\omega) \quad (\text{S.14})$$

$$+ \int R^2(H\omega) d\nu(\omega) \Big). \quad (\text{S.15})$$

We have split the integral formally. Using the bound $R(\omega) \leq C \|\omega\|_\infty^4$ (Equation (S.6)) we bound the integrands of Expressions (S.12), (S.13), (S.14), and (S.15). For some F we have

$$|(\omega^T H^T \Sigma_K H \omega)^2| \leq F \|\omega\|_\infty^4 \|H\|_\infty^4,$$

$$|R(H\omega) (\omega^T H^T \Sigma_K H \omega)| \leq F \|\omega\|_\infty^6 \|H\|_\infty^6,$$

$$|R(H\omega)| \leq F \|\omega\|_\infty^4 \|H\|_\infty^4,$$

$$|R^2(H\omega)| \leq F \|\omega\|_\infty^8 \|H\|_\infty^8.$$

Note that by the definition of $d\nu(\omega)$ and the fact $\int \|\omega\|_\infty^8 |f_\epsilon(\omega)|^2 d\omega < \infty$ (Assumption (9)) we have

$$\int \|\omega\|_\infty^8 d\nu(\omega) = \int \|\omega\|_\infty^8 |\hat{f}_\epsilon(\omega)|^2 d\omega - \int \|\omega\|_\infty^8 |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)|^2 d\omega < \infty.$$

So Expressions (S.12), (S.13), (S.14), and (S.15) are all integrable and $O(\|H\|_\infty^4)$. Thus

$$\frac{1}{n} \int |\hat{K}(H\omega)|^2 d\nu(\omega) = \frac{1}{n} \int d\nu(\omega) - \left(\frac{1}{n} \int (\omega^T H^T \Sigma_K H \omega) d\nu(\omega) \right) (1 + O(\|H\|_\infty^2)).$$

□

S.1.3 Proof of Theorem 3

Recall that $K = \phi_{\Sigma_K}$, $f_\epsilon = \phi_{\Sigma_\epsilon}$, and

$$f_X(x) = \sum_{j=1}^m \alpha_j \phi_{\Sigma_j}(x - \mu_j). \quad (\text{S.16})$$

Let $\alpha = (\alpha_1, \dots, \alpha_m)$, $S = H^T \Sigma_K H$, and Ω_a for $a \in \{0, 1, 2\}$ be a $m \times m$ matrix with j, j' entry equal to

$$\phi_{aS+2\Sigma_\epsilon+\Sigma_j+\Sigma_{j'}}(\mu_j - \mu_{j'}). \quad (\text{S.17})$$

In **Variance** we show

$$\int \text{Var}(\tilde{f}_{Y,H}(y)) dy = \frac{1}{n} (\phi_{2S+2\Sigma_\epsilon}(0) - \alpha^T \Omega_2 \alpha). \quad (\text{S.18})$$

In **Bias** we show

$$\int \left(\mathbb{E}[\tilde{f}_{Y,H}(y)] - f_Y(y) \right)^2 dy = \alpha^T (\Omega_2 - 2\Omega_1 + \Omega_0) \alpha. \quad (\text{S.19})$$

By the bias–variance decomposition of the MISE, we can sum Equation (S.18) and (S.19) to obtain the result

$$\text{MISE}(H) = \frac{1}{n} \phi_{2S+2\Sigma_\epsilon}(0) + \alpha^T ((1 - n^{-1})\Omega_2 - 2\Omega_1 + \Omega_0) \alpha.$$

First note that since ϵ and K_H are both normal, the estimator $\tilde{f}_{Y,H}$ has a simple form, specifically

$$\tilde{f}_{Y,H}(y) = \frac{1}{n} \sum_{i=1}^n \phi_{S+\Sigma_\epsilon}(y - X_i). \quad (\text{S.20})$$

Second note that (see e.g., A.2 on p.527 in Wand and Jones [1993]) for any two covariance matrices Σ and Σ' and mean vectors μ and μ' we have

$$\int \phi_\Sigma(x - \mu) \phi_{\Sigma'}(x - \mu') dx = \phi_{\Sigma+\Sigma'}(\mu - \mu'). \quad (\text{S.21})$$

Variance: Using Equation (S.20), we have

$$\int \text{Var}(\tilde{f}_{Y,H}(y)) dy = \frac{1}{n} \left(\int \mathbb{E}[\phi_{S+\Sigma_\epsilon}^2(y - X_1)] dy - \int \mathbb{E}[\phi_{S+\Sigma_\epsilon}(y - X_1)]^2 dy \right). \quad (\text{S.22})$$

We now simplify each term in the parenthesis on the right hand side of this equation. Using Equation (S.21) for the last equality, for the first term on the right hand side of Equation (S.22) we have

$$\begin{aligned} \int \mathbb{E}[\phi_{S+\Sigma_\epsilon}^2(y - X_1)] dy &= \int \int \phi_{S+\Sigma_\epsilon}^2(y - x) f_X(x) dx dy \\ &= \int \phi_{S+\Sigma_\epsilon}^2(y) dy \\ &= \phi_{2S+2\Sigma_\epsilon}(0). \end{aligned}$$

Recalling the representation of f_X in Equation (S.16), the identity in Equation (S.21), and the definition of Ω_2 in Equation (S.17), for the second term on the right hand side of Equation (S.22) we have

$$\begin{aligned} \int \mathbb{E}[\phi_{S+\Sigma_\epsilon}(y - X_1)]^2 dy &= \int \left(\int \phi_{S+\Sigma_\epsilon}(y - x) f_X(x) dx \right)^2 dy \\ &= \int \left(\int \phi_{S+\Sigma_\epsilon}(y - x) \left(\sum_{j=1}^m \alpha_j \phi_{\Sigma_j}(x - \mu_j) \right) dx \right)^2 dy \\ &= \int \left(\sum_{j=1}^m \alpha_j \phi_{S+\Sigma_\epsilon+\Sigma_j}(y - \mu_j) \right)^2 dy \\ &= \sum_{j=1}^m \sum_{j'=1}^m \alpha_j \alpha_{j'} \int \phi_{S+\Sigma_\epsilon+\Sigma_j}(y - \mu_j) \phi_{S+\Sigma_\epsilon+\Sigma_{j'}}(y - \mu_{j'}) dy \\ &= \sum_{j=1}^m \sum_{j'=1}^m \alpha_j \alpha_{j'} \phi_{2S+2\Sigma_\epsilon+\Sigma_j+\Sigma_{j'}}(\mu_j - \mu_{j'}) \\ &= \alpha^T \Omega_2 \alpha. \end{aligned}$$

Hence

$$\int \text{Var}(\tilde{f}_{Y,H}(y)) dy = \frac{1}{n} (\phi_{2S+2\Sigma_\epsilon}(0) - \alpha^T \Omega_2 \alpha)$$

Bias: Recalling $f_\epsilon = \phi_{\Sigma_\epsilon}$ and the representations of f_X and $\tilde{f}_{Y,H}$ in Equations (S.16) and (S.20), note

$$\begin{aligned} f_Y(y) &= \int f_X(y - \epsilon) f_\epsilon(\epsilon) d\epsilon = \sum_{j=1}^m \alpha_j \int \phi_{\Sigma_j}(y - \epsilon - \mu_j) \phi_{\Sigma_\epsilon}(\epsilon) d\epsilon = \sum_{j=1}^m \alpha_j \phi_{\Sigma_j + \Sigma_\epsilon}(y - \mu_j), \\ \mathbb{E}[\tilde{f}_{Y,H}(y)] &= \int \phi_{S+\Sigma_\epsilon}(y - x) \sum_{j=1}^m \alpha_j \phi_{\Sigma_j}(x - \mu_j) dx = \sum_{j=1}^m \alpha_j \phi_{S+\Sigma_j+\Sigma_\epsilon}(y - \mu_j). \end{aligned}$$

Using these identities and the definition of Ω_a (Equation (S.17)), for the integrated squared bias we have

$$\begin{aligned} \int \left(\mathbb{E}[\tilde{f}_{Y,H}(y)] - f_Y(y) \right)^2 dy &= \int \left(\sum_{j=1}^m \alpha_j (\phi_{S+\Sigma_j+\Sigma_\epsilon}(y - \mu_j) - \phi_{\Sigma_j+\Sigma_\epsilon}(y - \mu_j)) \right)^2 dy \\ &= \int \sum_{j=1}^m \sum_{j'=1}^m \alpha_j \alpha_{j'} \left(\phi_{S+\Sigma_j+\Sigma_\epsilon}(y - \mu_j) \phi_{S+\Sigma_{j'}+\Sigma_\epsilon}(y - \mu_{j'}) \right. \\ &\quad \left. - \phi_{S+\Sigma_j+\Sigma_\epsilon}(y - \mu_j) \phi_{\Sigma_{j'}+\Sigma_\epsilon}(y - \mu_{j'}) \right. \\ &\quad \left. - \phi_{S+\Sigma_{j'}+\Sigma_\epsilon}(y - \mu_{j'}) \phi_{\Sigma_j+\Sigma_\epsilon}(y - \mu_j) \right. \\ &\quad \left. + \phi_{\Sigma_j+\Sigma_\epsilon}(y - \mu_j) \phi_{\Sigma_{j'}+\Sigma_\epsilon}(y - \mu_{j'}) \right) dy \\ &= \sum_{j=1}^m \sum_{j'=1}^m \alpha_j \alpha_{j'} \left(\phi_{2S+2\Sigma_\epsilon+\Sigma_j+\Sigma_{j'}}(\mu_j - \mu_{j'}) \right. \\ &\quad \left. - 2\phi_{S+2\Sigma_\epsilon+\Sigma_j+\Sigma_{j'}}(\mu_j - \mu_{j'}) + \phi_{2\Sigma_\epsilon+\Sigma_j+\Sigma_{j'}}(\mu_j - \mu_{j'}) \right) \\ &= \alpha^T (\Omega_2 - 2\Omega_1 + \Omega_0) \alpha. \end{aligned}$$

S.1.4 Lemmas

Lemma 1. *Under Assumptions A, K is a symmetric density function in \mathbb{R}^p with a characteristic function \hat{K} that is four times continuously differentiable. Let Σ_K be the variance of K . We Taylor expand \hat{K} around 0, obtaining*

$$\hat{K}(\omega) = 1 - \frac{\omega^T \Sigma_K \omega}{2} + R(\omega).$$

There exists C such that for any ω

$$R(\omega) \leq C \|\omega\|_\infty^4.$$

Proof. We bound the remainder term $R(\omega)$ by considering two cases.

1. $\{\omega : \|\omega\|_\infty \leq 1\}$: Since \hat{K} is four times continuously differentiable, there exists D such that for any $\{j : \sum_{k=1}^p j_k = 4\}$, $\forall \|\omega\|_\infty \leq 1$

$$\frac{\partial^4 \hat{K}}{\partial \omega_1^{j_1} \dots \partial \omega_p^{j_p}}(\omega) < D. \quad (\text{S.23})$$

Using the mean value form of the Taylor remainder we have (see e.g. Theorem 7.1 in Edwards Jr [1973] on page 131)

$$R(\omega) = \sum_{\{j: \sum_{k=1}^p j_k = 4\}} \frac{\partial^4 \widehat{K}}{\partial \omega_1^{j_1}, \dots, \partial \omega_p^{j_p}}(\xi) \prod_{k=1}^p \frac{\omega_k^{j_k}}{j_k!}.$$

for some $\xi = t\omega$ for $t \in [0, 1]$. Using Equation (S.23) and noting $\prod_{k=1}^p \omega_k^{j_k} \leq \|\omega\|_\infty^4$, for some C we have

$$|R(\omega)| \leq C \|\omega\|_\infty^4.$$

2. $\{\omega : \|\omega\|_\infty > 1\}$: Note that for some D , $\frac{\omega^T \Sigma_K \omega}{2} \leq D \|\omega\|_\infty^2$. Also note that on the set $\|\omega\|_\infty > 1$ we have $\|\omega\|_\infty^2 \leq \|\omega\|_\infty^4$. We have

$$\begin{aligned} |R(\omega)| &= \left| \widehat{K}(\omega) - 1 + \frac{\omega^T \Sigma_K \omega}{2} \right| \\ &\leq |\widehat{K}(\omega)| + |1| + \left| \frac{\omega^T \Sigma_K \omega}{2} \right| \\ &\leq 2 + \left| \frac{\omega^T \Sigma_K \omega}{2} \right| \\ &\leq 2 + D \|\omega\|_\infty^2 \\ &\leq 2 \|\omega\|_\infty^2 + D \|\omega\|_\infty^2 \\ &\leq (2 + D) \|\omega\|_\infty^4 \end{aligned}$$

□

S.2 Technical Notes

S.2.1 Full Bandwidth Matrix Optimization

In Theorem 2 on p.7, the MISE (using a full bandwidth matrix) is

$$\frac{1}{n} \int d\nu(\omega) + \left(\frac{1}{4} \int (\omega^T S \omega)^2 d\mu(\omega) - \frac{1}{n} \int (\omega^T S \omega) d\nu(\omega) \right) (1 + O(\|H\|_\infty^2))$$

where $S = H^T \Sigma_K H$. Using vec notation and the identity $\text{vec}(EFG) = (G^T \otimes E) \text{vec}(F)$ where \otimes denotes Kronecker product (see Equation 5 on page 67 in Henderson and Searle [1979]), we write the optimization problem for S as

$$S_Y^* = \underset{S \succeq 0}{\text{argmin}} \text{vec}(S)^T B \text{vec}(S) - \frac{1}{n} \text{vec}(S)^T V \quad (\text{S.24})$$

where

$$\begin{aligned} B &= \frac{1}{4} \int (\omega \otimes \omega)(\omega \otimes \omega)^T d\mu(\omega), \\ V &= \int (\omega \otimes \omega) d\nu(\omega). \end{aligned}$$

It is important to note that B and V cannot be computed from the data because they depend on the unknown function $\hat{f}_X(\omega)$. In practice we could use plug-in estimators to approximate these integrals.

The unconstrained solution to optimization problem (S.24) may not be positive semidefinite, so we cannot omit the $S \succeq 0$ constraint and use a quadratic solver (see Subsection S.2.2 for an example). Also note that one cannot analytically solve the unconstrained version of optimization problem (S.24) and then check whether the resulting S_Y^* is positive semidefinite. In other words, the following procedure is not valid:

$$\begin{aligned} g(\text{vec}(S)) &\equiv \text{vec}(S)^T B \text{vec}(S) - \frac{1}{n} \text{vec}(S)^T V, \\ \implies \nabla g(\text{vec}(S)) &= 2B \text{vec}(S) - \frac{1}{n} V. \end{aligned}$$

Setting the gradient equal to 0 and solving we have

$$\text{vec}(S_Y^*) = \frac{1}{2n} B^{-1} V.$$

One could then check whether $S_Y^* \succeq 0$. This procedure is not valid because B is not invertible. To see that B is not invertible, note that the vector $(\omega \otimes \omega)$ has p^2 elements, but not p^2 unique elements. For example when $p = 2$, $(\omega \otimes \omega) = (\omega_1^4, \omega_1 \omega_2, \omega_1 \omega_2, \omega_2^2)^T$. When the j th and k th elements of $(\omega \otimes \omega)$ are equal, the j th and k th rows of $(\omega \otimes \omega)(\omega \otimes \omega)^T$ are equal. Thus at least two rows of $B \equiv \int (\omega \otimes \omega)(\omega \otimes \omega)^T d\mu(\omega)$ are equal, implying that B cannot be inverted.

S.2.2 Diagonal Bandwidth and $\Sigma_K = I$

In Equation (11) the full bandwidth matrix asymptotic expansion of the MISE was presented. By restricting the kernel to have $\Sigma_K = I$ and the bandwidth matrix to be diagonal we achieve considerable simplification of the MISE. Let $h_i = H_{ii}$ and $s = (h_1^2, \dots, h_p^2)$. The MISE (Equation (11)) becomes

$$(2\pi)^p \text{MISE}(s) = \frac{1}{n} \int d\nu(\omega) + \left(s^T B s - \frac{1}{n} s^T V \right) (1 + O(\|s\|_\infty)),$$

where

$$\begin{aligned} B_{i,j} &= \frac{1}{4} \int \omega_i^2 \omega_j^2 d\mu(\omega), \\ V_i &= \int \omega_i^2 d\nu(\omega). \end{aligned}$$

We seek the s which minimizes the larger order terms in the MISE expression. In other words we seek

$$s_Y^* = \underset{s \geq 0}{\text{argmin}} \left(s^T B s - \frac{1}{n} s^T V \right). \quad (\text{S.25})$$

B is positive definite so the expression is strictly convex and there is a unique solution. Enforcing the domain restriction $s \geq 0$ (elementwise) is necessary: even in simple cases, the unconstrained optimum $\frac{1}{2n} B^{-1} V$ may have elements less than 0. In the following paragraphs we work through an example where f_X and f_ϵ are bivariate independent normals with ϵ having small variance along one direction. The kernel is normal with identity covariance. The normality is not essential for this example, but makes the computations simpler.

We begin by showing that the optimal bandwidth matrix is diagonal, implying that optimizing over the full bandwidth matrix and the diagonal matrix are equivalent. We then show that when optimizing over the unconstrained diagonal matrix, the direction in which ϵ has larger variance yields a “negative squared bandwidth”. Consider:

$$\begin{aligned} f_X &\sim N(0, I_{2 \times 2}), \\ f_\epsilon &\sim N(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}), \\ \Sigma_K &\equiv \int xx^T K(x) dx = I_2 \times 2. \end{aligned}$$

We parameterize the bandwidth matrix using $H = \begin{bmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{bmatrix}$. First consider optimizing over the entire bandwidth matrix, Equation (S.24). In our case

$$\begin{aligned} S &\equiv H^T \Sigma_K H = H^T H, \\ B &= \int (\omega \otimes \omega)(\omega \otimes \omega)^T d\mu(\omega) = \int \begin{bmatrix} \omega_1^4 & \omega_1^3 \omega_2 & \omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 \\ \omega_1^3 \omega_2 & \omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 & \omega_2^4 \\ \omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 & \omega_2^4 & \omega_1^3 \omega_2 \\ \omega_1 \omega_2^3 & \omega_2^4 & \omega_1^3 \omega_2 & \omega_1^2 \omega_2^2 \end{bmatrix} d\mu(\omega), \\ V &= \int (\omega \otimes \omega) d\nu(\omega) = \int \begin{bmatrix} \omega_1^2 \\ \omega_1 \omega_2 \\ \omega_1 \omega_2 \\ \omega_2^2 \end{bmatrix} d\nu(\omega). \end{aligned}$$

So Equation (S.24) becomes

$$\begin{aligned} &\text{vec}(H^T H)^T \int \begin{bmatrix} \omega_1^4 & \omega_1^3 \omega_2 & \omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 \\ \omega_1^3 \omega_2 & \omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 & \omega_2^4 \\ \omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 & \omega_2^4 & \omega_1^3 \omega_2 \\ \omega_1 \omega_2^3 & \omega_2^4 & \omega_1^3 \omega_2 & \omega_1^2 \omega_2^2 \end{bmatrix} d\mu(\omega) \text{vec}(H^T H) \\ &- \frac{1}{n} \text{vec}(H^T H)^T \left(\int \begin{bmatrix} \omega_1^2 \\ \omega_1 \omega_2 \\ \omega_1 \omega_2 \\ \omega_2^2 \end{bmatrix} d\nu(\omega) \right). \end{aligned}$$

The integration causes those terms involving odd powers of ω_i to be 0 by independence and symmetry of $d\nu(\omega)$ and $d\mu(\omega)$. Additionally the center $\omega_1^2 \omega_2^2$ terms are moved outside the main expression and into the

third term. We have

$$\begin{aligned} & \text{vec}(H^T H)^T \int \begin{bmatrix} \omega_1^4 & 0 & 0 & \omega_1^2 \omega_2^2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \omega_1^2 \omega_2^2 & 0 & 0 & \omega_2^4 \end{bmatrix} d\mu(\omega) \text{vec}(H^T H) \\ & - \frac{1}{n} \text{vec}(H^T H)^T \left(\int \begin{bmatrix} \omega_1^2 \\ 0 \\ 0 \\ \omega_2^2 \end{bmatrix} d\nu(\omega) \right) + 4(h_{12}(h_{11} + h_{22}))^2 \int \omega_1^2 \omega_2^2 d\mu(\omega). \end{aligned}$$

Since

$$H^T H = \begin{bmatrix} h_{11}^2 + h_{12}^2 & h_{12}(h_{11} + h_{22}) \\ h_{12}(h_{11} + h_{22}) & h_{12}^2 + h_{22}^2 \end{bmatrix},$$

minimization of the first two terms depends on $(h_{11}^2 + h_{12}^2, h_{22}^2 + h_{12}^2)$. So by setting $h_{12} = 0$ we make the third term in the expression 0, without restricting minimization of the first two terms. Thus for the general bandwidth matrix the minimum occurs when the off-diagonal elements are 0.

Now let $s = (h_{11}^2, h_{22}^2)$. We study the diagonal optimization problem (S.25)

$$s_Y^* = \min_s s^T B' s - \frac{1}{n} s^T V',$$

where

$$\begin{aligned} B'_{i,j} &= \frac{1}{4} \int \omega_i^2 \omega_j^2 d\mu(\omega) = \frac{1}{4} \int \omega_i^2 \omega_j^2 |\widehat{f}_X(\omega)|^2 |\widehat{f}_\epsilon(\omega)|^2 d\omega, \\ V'_i &= \int \omega_i^2 d\nu(\omega) = \int \omega_i^2 |\widehat{f}_\epsilon(\omega)|^2 d\omega - \int \omega_i^2 |\widehat{f}_X(\omega)|^2 |\widehat{f}_\epsilon(\omega)|^2 d\omega. \end{aligned}$$

With no restrictions on s the optimum is

$$s_Y^* = \frac{1}{2n} B'^{-1} V'.$$

We now compute this quantity for the given densities. First compute B' :

$$\begin{aligned} 4B'_{11} &= \int \omega_1^4 |\widehat{f}_{X_1}(\omega_1)|^2 |\widehat{f}_{\epsilon_1}(\omega_1)|^2 d\omega_1 \int |\widehat{f}_{X_2}(\omega_2)|^2 |\widehat{f}_{\epsilon_2}(\omega_2)|^2 d\omega_2 \\ &= \left(\frac{3}{4} \sqrt{\frac{\pi}{(1 + \sigma_1^2)^5}} \right) \left(\sqrt{\frac{\pi}{1 + \sigma_2^2}} \right), \\ 4B'_{22} &= \int \omega_2^4 |\widehat{f}_{X_2}(\omega_2)|^2 |\widehat{f}_{\epsilon_2}(\omega_2)|^2 d\omega_2 \int |\widehat{f}_{X_1}(\omega_1)|^2 |\widehat{f}_{\epsilon_1}(\omega_1)|^2 d\omega_1 \\ &= \left(\frac{3}{4} \sqrt{\frac{\pi}{(1 + \sigma_2^2)^5}} \right) \left(\sqrt{\frac{\pi}{1 + \sigma_1^2}} \right), \\ 4B'_{12} &= \int \omega_1^2 |\widehat{f}_{X_1}(\omega_1)|^2 |\widehat{f}_{\epsilon_1}(\omega_1)|^2 d\omega_1 \int \omega_2^2 |\widehat{f}_{X_2}(\omega_2)|^2 |\widehat{f}_{\epsilon_2}(\omega_2)|^2 d\omega_2 \\ &= \left(\frac{1}{2} \sqrt{\frac{\pi}{(1 + \sigma_1^2)^3}} \right) \left(\frac{1}{2} \sqrt{\frac{\pi}{(1 + \sigma_2^2)^3}} \right). \end{aligned}$$

Since B' and B'^{-1} are symmetric, we write only the upper triangle:

$$B' = \frac{\pi}{16} \begin{bmatrix} 3\frac{1}{\sqrt{(1+\sigma_1^2)^5(1+\sigma_2^2)}} & \frac{1}{\sqrt{(1+\sigma_1^2)^3(1+\sigma_2^2)^3}} \\ & 3\frac{1}{\sqrt{(1+\sigma_2^2)^5(1+\sigma_1^2)}} \end{bmatrix}.$$

Taking the inverse we obtain

$$\begin{aligned} B'^{-1} &= \frac{2(1+\sigma_1^2)^3(1+\sigma_2^2)^3}{\pi} \begin{bmatrix} 3\frac{1}{\sqrt{(1+\sigma_2^2)^5(1+\sigma_1^2)}} & -\frac{1}{\sqrt{(1+\sigma_1^2)^3(1+\sigma_2^2)^3}} \\ & 3\frac{1}{\sqrt{(1+\sigma_1^2)^5(1+\sigma_2^2)}} \end{bmatrix} \\ &= \frac{2}{\pi} \begin{bmatrix} 3\sqrt{(1+\sigma_2^2)(1+\sigma_1^2)^5} & -\sqrt{(1+\sigma_1^2)^3(1+\sigma_2^2)^3} \\ & 3\sqrt{(1+\sigma_1^2)(1+\sigma_2^2)^5} \end{bmatrix}. \end{aligned}$$

For V' we have

$$\begin{aligned} V' &= \frac{\pi}{2} \left(\begin{bmatrix} \sigma_1^{-3}\sigma_2^{-1} \\ \sigma_1^{-1}\sigma_2^{-3} \end{bmatrix} - \begin{bmatrix} \frac{1}{\sqrt{(1+\sigma_1^2)^3(1+\sigma_2^2)}} \\ \frac{1}{\sqrt{(1+\sigma_2^2)^3(1+\sigma_1^2)}} \end{bmatrix} \right) \\ &= \frac{\pi}{2\sigma_2^3} \left(\begin{bmatrix} 0 \\ \sigma_1^{-1} \end{bmatrix} + \sigma_2^2 \begin{bmatrix} \sigma_1^{-3} \\ 0 \end{bmatrix} - \sigma_2^3 \begin{bmatrix} \frac{1}{\sqrt{(1+\sigma_1^2)^3(1+\sigma_2^2)}} \\ \frac{1}{\sqrt{(1+\sigma_2^2)^3(1+\sigma_1^2)}} \end{bmatrix} \right). \end{aligned}$$

So the optimal s is

$$\begin{aligned} s_Y^* &= \frac{1}{2n} B'^{-1} V' \\ &= \frac{1}{2n\sigma_2^3} \left(\begin{bmatrix} -\sigma_1^{-1}\sqrt{(1+\sigma_1^2)^3(1+\sigma_2^2)^3} \\ 3\sigma_1^{-1}\sqrt{(1+\sigma_1^2)(1+\sigma_2^2)^5} \end{bmatrix} \right. \\ &\quad \left. + \sigma_2^2 \begin{bmatrix} 3\sigma_1^{-3}\sqrt{(1+\sigma_2^2)(1+\sigma_1^2)^5} \\ -\sigma_1^{-3}\sqrt{(1+\sigma_1^2)^3(1+\sigma_2^2)^3} \end{bmatrix} - 2\sigma_2^3 \begin{bmatrix} 1+\sigma_1^2 \\ 1+\sigma_2^2 \end{bmatrix} \right). \end{aligned}$$

For σ_2 close to 0 and small relative to σ_1 this quantity is approximately

$$s_Y^* \approx \frac{1}{2n\sigma_1\sigma_2^3} \left(\begin{bmatrix} -\sqrt{(1+\sigma_1^2)^3} \\ 3\sqrt{(1+\sigma_1^2)} \end{bmatrix} \right). \quad (\text{S.26})$$

The unconstrained optimization results in an s_Y^* with negative elements.